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

Analysis of Chinese Syllable Durations in Running Speech of Japanese L2 Learners

Yue Sun[1], *Shudon Hsiao*[1], *Yoshinori Sagisaka*[1], *Jinsong Zhang*[2]

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Japan
[2]College of Information Science, Beijing Language and Culture University, China

yue.cherry.sun@gmail.com, xudong.ruri@gmail.com
ysagisaka@gmail.com, jinsong.zhang@blcu.edu.cn

Abstract

Aiming at better understanding of prosody generation by native Japanese learners of Mandarin as a second language (L2), we analyzed the syllable duration differences between tone types. By comparing the mean syllable durations and the variation of normalized syllable durations across tone types and speakers, significant differences were found between tone types as well as between speakers. Native Chinese speakers generate tone 1 and tone 2 with relatively long durations but smaller variations, contrary to tone 3 and tone 4. Japanese L2 learners generate tone 3 with relatively high variations compared to the other tones, while the mean duration of tone 4 was remarkably different from natives. Compared with native speakers, the variations of both tone 3 and tone 4 are significantly smaller. Furthermore, the neutral tone caused a significant increase of the mean variation across tones for the Japanese L2 learners. The results suggest that native Chinese speakers control syllable durations adaptively with tones, especially for tone 3 and tone 4, in running speech while Japanese L2 learners tend to pronounce them in isolated syllable fashion.

Index Terms: syllable duration, Mandarin Chinese tones, Japanese L2 learners

1. Introduction

Mandarin Chinese is a typical tone language which consists of four lexical tones and one neutral tone. They have great importance for teaching Chinese as a second language (L2). The most remarkable acoustic features of these tone differences are found in the fundamental frequency (F0). Depending on the F0 contours, the citation forms of the four lexical tones were separated to high level (T1), rising (T2), low-dipping/falling-rising (T3) and falling (T4). The neutral tone (T0) in Mandarin Chinese is the tone occurring in unstressed syllables, which do not have specific pitch value of their own but vary according to the tone of the preceding syllable [1].

The tonal information of a syllable in Mandarin has an influence on the syllable duration. It is well known that Mandarin Chinese tones have intrinsic durations when they are pronounced in isolated monosyllables. T2 and T3 tend to be the longest, T4 the shortest [2]. Generally the neutral tone syllable shows shorter duration and/or lower intensity than tonic syllables, and the duration was reported as the most important cue for identifying it [3].

There is evidence that the duration of tones varies in running speech. From syllable analysis using large speech databases, it has been reported that in increasing order, the mean syllable durations were T0, T3, T4, T2, T1 [4]. The possible reason is that tones are not presented individually in running speech, and hence that there is interaction between intonation and lexical tones [5]. When a tonal syllable is at the initial/final position or at the focus position of a sentence, it is obviously longer than at the unfocused medial position [6] [7]. In addition, T3 is uttered as a half third tone (it falls but does not rise) when it is followed by any tone except another T3 in which case tone sandhi occurs.

On the other hand, tone acquisition has always been a difficult task for second language learners of Mandarin Chinese, especially for those whose native languages are non-tonal. For Japanese L2 learners, the confusion of T2 and T3 in isolated syllables and bi-syllablesare already very well understood [8][9] , while their tone generation capability in running speech needs to be investigated further.

Since prosody generation is not only facilitated through different F0 patterns but also through different durational patterns for syllables, the existence of a dependency between tonal information, syllable duration and prosody generation can be expected. Furthermore, to realize the specific prosodic structure of a given sentence, the speaker has to control the vocal fold to generate the F0 information simultaneously with the durational information. As the capacity to control one's vocal fold for the realization of Mandarin Chines prosody can be expected to be different between native Mandarin speakers and L2 learners, different F0 patterns as well as different durational patterns for native speakers and L2 learners should also be expected.

Therefore, from the perspective of understanding prosody generation by L2 learners of Mandarin Chinese, we analyzed the tone effects on syllable durations for native Mandarin Chinese speakers and Japanese L2 learners in the present study. From the difference of syllable durations depending on tonal information between native Mandarin speakers and Japanese L2 learners, the learner's tone generation capability could also be revealed.

The remainder of the paper is organized as follows. After introducing the speech corpus and the data collection in the following Section 2, we compare the syllable duration differences of tone types between native speech and learner's speech by calculating the mean syllable durations in Section 3, and by calculating the variations of the normalized syllable durations in Section 4. Afterwards, we summarize the results and suggest future work in the final Section 5.

2. Data collection

41 speakers participated in the speech corpus collection. 12 of them (6 female and 6 male) were native Mandarin Chinese

http://dx.doi.org/10.21437/Interspeech.2016-824

speakers. They were residents in Beijing and had reached the 1st level of PSC (the official language test of Mandarin for native speakers of Chinese). 19 (9 female and 10 male) were Intermediate Japanese L2 learners. They were students of Mandarin Intensive classes at Beijing Language and Culture University (BLCU) in China. Those two groups of speakers were participants for the collection of the BLCU inter-Chinese speech corpus. 10 (7 female and 3 male) were Japanese L2 learners at the beginner's level who had been studying Chinese for one year in a regular Mandarin class at Waseda University, Japan. In this study, we will refer to the three groups of speakers as native/Chinese, intermediate learner and beginner respectively.

50 spoken language sentences were selected to the corpus-collection. The sampling frequency for recording was 44100Hz, 16bit. In the end, since a few students failed to participate in parts of the data collection, there were average 49 running speech samples for each native speaker, 44 for each intermediate learner, 51 for each beginner (including a self introduction speech for each speaker). The speech samples consist of 9.6 syllables on average.

All the speech samples were manually annotated, each syllable with a tone was labeled as an interval, and mispronounced tones were marked. The open source Praat tool ProsodyPro(Version 5.5.6) was used to extract the durational information from the annotated speech samples [10]. Only the syllables without tone mispronunciation were used for the data analysis. Research about the third-tone sandhi (the first T3 of the T3T3 combination becomes T2) indicates that the F0 contours of the third-tone sandhi T2 and the regular T2 are statistically equal [11], and that native Mandarin speakers could not differentiate third-tone sandhi T2 and regular T2 perceptually [12]. Thus, in this study, although the third-tone sandhi T2 syllables were separated during the annotation, they were merged with the regular T2 syllables for the data analysis.

3. Tone effect on mean syllable duration

The syllable duration varies by tone types for native speakers as well as for Japanese L2 learners. Figure 1. shows the mean syllable durations in millisecond of tones for each speaker group.

By ranking the mean syllable duration of tones increasingly, a clear difference for T4 between the native speakers and the Japanese L2 learners can be observed. The increasing order

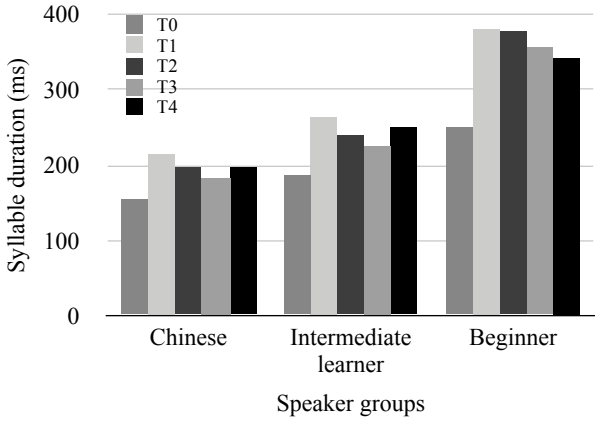

Figure 1: *The mean syllable duration of each tone for native Chinese speakers, Intermediate Japanese L2 learners and Japanese L2 beginners.*

for the natives is T0, T3, T4, T2, T1, which is identical to the result of the large corpus based study [4]. It is T0, T3, T2, T4, T1 for the intermediate learners, as they tended to produce T4 somewhat longer than T2. For the beginners an order of T0, T4, T3, T2, T1 can be observed, as they apparently uttered T4 longer than T0 but shorter than the other tones.

As the syllable duration for T1 is the longest and T0 is the shortest for all speaker groups, the maximal variations in terms of average syllable durations is given by the durational differences between T1 and T0, which is 58ms for the natives, 77ms for the intermediate learners and 131ms for the beginners, respectively. Although the durational differences between T1 and T0 syllables are large, the one between the four lexical tone are much smaller. Subtracting the average syllable duration of the shortest lexical tone from the longest one, we get 31ms for natives, 38ms for intermediate learners and 39ms for beginners. The syllable duration of T0 clearly influenced the tone differences for Japanese L2 learners.

Another issue should be taken into consideration for the further study. Figure 1 also showed that the speaking rates are obviously slower for the Japanese L2 learners, especially for the beginners. By taking average over all tones, the mean syllable duration was 191ms for natives, 237ms for intermediate learners and 341ms for beginners. It could have an impact on the result that we compared the difference between speaker groups using data in millisecond unit. Therefore, a normalization method has been carried out before the comparison study.

4. Tone effect on syllable duration variance

4.1. Variation of normalized syllable duration

To remove the influence of speaking rate, normalization was carried out for each speaker. It alternates the unit of syllable duration from millisecond to a normalized scale. The variations of normalized syllable duration for each speaker as well as for each tone type are calculated in the normalized scale given through equation (1).

$$V_{norm} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{dur_i}{\overline{dur}} - 1\right)^2 \quad (1)$$

Here for a specific speaker or a specific tone type, n is the total amount of syllables, dur_i is the duration in ms of the i-th syllable and $\overline{dur}$ is the mean duration of the syllables.

4.2. Syllable duration effect of neutral tone

By looking at the syllable duration of each tone in millisecond, T0 was obviously shorter than other tones especially for beginners. It is reasonable to assume that the short syllable duration for T0 increases the variation of syllable durations. Therefore, if the data of T0 was excluded, the variation value would decrease, especially for the beginners. To validate this assumption, the mean variations of normalized syllable duration, when including and when excluding T0 data, were compared for each speaker group.

Figure 2 shows the mean variations of normalized syllable durations for each speaker group, where one bar refers to the analysis including T0 and one to the analysis excluding it, respectively. Whether including the data for T0 or not, natives had the largest variations while beginners had the smallest variations among the three speaker groups. Not only for Japanese L2 learners, but also for natives, the variations did decrease after excluding T0 data. The difference between including and

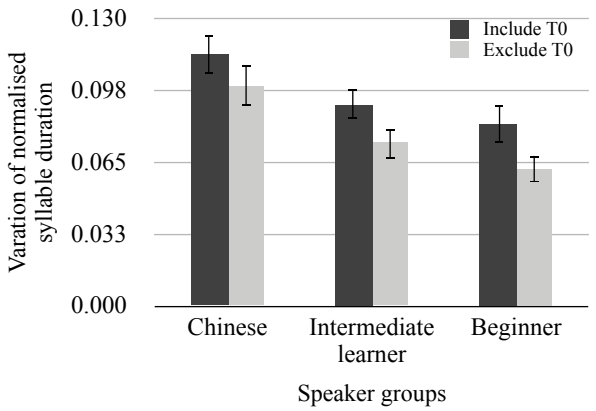

Figure 2: *The variation of normalized syllable duration when including and excluding the T0 data for native Chinese speakers, Intermediate Japanese L2 learners and Japanese L2 beginners.*

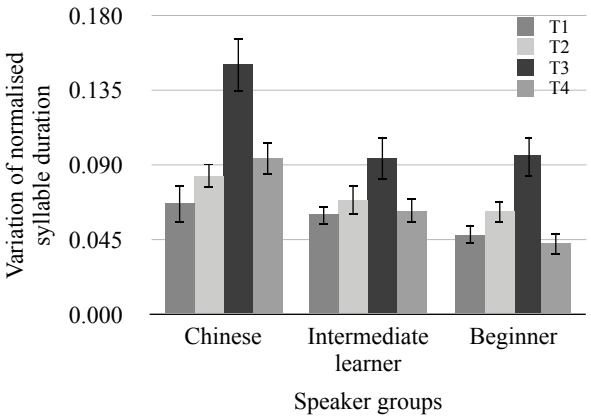

Figure 3: *The variations of normalized syllable durations of the four lexical tones for native Chinese speakers, Intermediate Japanese L2 learners and Japanese L2 beginners.*

Table 1: *Significance between speaker groups when including or excluding T0 from the dataset.*

	Include T0	Exclude T0
Intermediate learner - Chinese	0.000***	0.000***
Beginner - Chinese	0.000***	0.000***
Intermediate learner - Beginner	0.039	0.018

excluding T0 data was the smallest for natives and the largest for beginners.

Statistical analyses were carried out by using a 3×2 (speaker groups × include/exclude T0 data) ANOVA. There was a main effect of speaker groups [$F(2,38)=125.395$, $p<0.001$], which showed that the variations were different between the three speaker groups. There was also a main effect of whether the T0 data was included or not [$F(1,39)=25.133$, $p<0.001$], where the variations were impaired by the T0 data.

Tukey's HSD tests were used to analyze the differences in detail. The difference between including and excluding the data for T0 was not significant for natives ($p=0.294$), slightly significant for beginners ($p=0.024$), and significant for intermediate learners ($p=0.006$). It means that the shorter duration of T0 clearly had impact on the durational differences for both beginners and intermediate learners, since they uttered the other tones significantly longer than T0. But although the variations decreased when excluding the T0 data, it did not influence the fact that the variations are higher for native Chinese speakers.

Table 1 lists the results of a Tukey's HSD test for comparing the differences between speaker groups when including and excluding the data for T0. The differences between natives and both the two Japanese L2 learner groups were significant ($p<0.01$), and the difference between the two Japanese L2 learner groups themselves was slightly significant ($p<0.05$). This indicates that whether or not T0 data was included, the variations of syllable durations were significantly different between native speakers and learners, although the language proficiency of the intermediate learners was higher than that of the beginners. The influence of the shortness of T0 was not the only reason for the durational differences between natives and learners which means that there are other significant durational variations between the speaker groups.

Table 2: *Significance between tone pairs for native Chinese speakers, Intermediate Japanese L2 learners and Japanese L2 beginners.*

	Chinese	Intermediate learner	Beginner
T2-T1	1.000	1.000	1.000
T3-T1	0.000***	0.000***	0.000***
T4-T1	0.000***	1.000	0.999
T3-T2	0.000***	0.000***	0.000***
T4-T2	0.001**	1.000	0.991
T4-T3	0.000***	0.000***	0.000***

4.3. Syllable duration effect of lexical tones

Except the short T0 syllables, the small variations of T3 and T4 syllable durations affected the mean variations for intermediate learners and beginners. Figure 3 shows the mean variations of normalized syllable durations of the four lexical tones for each speaker group. It clearly shows that, despite tone types, the variations for native Chinese speakers were larger than the ones for intermediate learners and beginners. The variation of T3 was particularly large for the natives, followed by T4, T2, and T1. For the two groups of Japanese L2 learners, the variations of T3 were larger than for the other tone types, but the values were much smaller than for the natives. Furthermore, T4 had the smallest variations among the four tones. Between those two Japanese L2 learner groups, the intermediate learners had larger variations than the beginners, which could be a result of language proficiency. The results indicate that native Chinese speakers utter the four lexical tones with highly variating syllable durations in running speech, especially for T3 syllables, while Japanese L2 learners produce the four tones with more similar syllable durations by reducing the variations of T3 and T4 syllables compared to native speakers.

A 3×4 (speaker groups × tones) ANOVA was conducted to analyze the differences between natives and Japanese L2 learners. There was a main effect of tone types [$F(3,37)=115.935$, $p<0.001$], which means that the syllable durations vary prominently between tones. There was a main effect of speaker groups [$F(2,38)=101.675$, $p<0.001$], where the variations were different between speaker groups.

Tukey's HSD tests were used to analyze the differences in

detail. The differences between speaker groups of each tone showed that, there was no difference between any speaker group for T1 or T2 ($p>0.05$). The differences of T3 were significant between the natives and both the intermediate and beginner Japanese L2 learner groups ($p<0.01$), but not significant between the two Japanese L2 learner groups themselves ($p>0.05$). The differences of T4 were significant between all the speaker groups ($p<0.01$). It confirmed the differences for T3 and T4 between natives and Japanese L2 learners shown by the mean variation of normalized duration, and it also indicates that the language proficiency has an impact on the tone production of T4 for Japanese L2 learners.

The differences between tone pairs for each speaker group are shown in Table 2. For the native Chinese speakers, except for th T1-T2 pair ($p>0.05$), the differences between the other tones were significant ($p<0.01$). For the two Japanese L2 learner groups, the differences were significant when comparing T3 with the other tones ($p<0.01$) but not significant for other tone pairs ($p>0.05$). This clearly revealed the fact that Japanese L2 learners reduce the durational variations of T4 syllables by producing it equally small variations as for T1 and T2.

5. Discussion

Combining the results, tone effects on syllable durations could clearly be seen for native Chinese speakers. Although the mean durations were similar for the four lexical tones, the variation for normalized syllable durations showed that they adapt the syllable durations of all the tones according to the context of the running speech. They produced T1 and T2 relatively longer and with more stable duration while T3 and T4 are shorter and more flexible in terms of duration. T3 varied the most in running speech. In this way the native Chinese speakers generated speech with high variations in terms of syllable duration even without taking the short T0 into account.

Japanese L2 learners showed different patterns for the relationship between tone type and syllable duration. The variations were mainly realized through a short T0 and partly through T3 while T1, T2 and T4 were produced with stable duration. Furthermore, the beginners showed more isochronous patterns than the intermediate learners.

The different patterns for Japanese L2 learners revealed their problems of controlling their vocal folds to generate tones according to Mandarin prosody structures. Because T3 could be implemented as either a low-dipping half-third tone, a rising tone or a complement falling-rising tone in running speech, it has the largest variation. As a consequence, Japanese L2 learners showed the biggest differences of T3 among the four lexical tones compared with native Chinese speakers.

Although teaching tones to L2 learners is a common issue in second language education of Mandarin Chinese, the phonetic instruction of tones is usually taking place at the beginning of the elementary courses, by focusing on the generation of tones in isolated fashion or bisyllabic fashion. The results of the present study provide evidence for the existence of tone generation problems for L2 learners in running speech where T3 causes the main problem. Future studies need to be conducted to explain the causes of durational variations of syllables by tone as well as the relative F0 alteration in native speech, thus to clarify the problem of learner's speech and eventually provide specific guidance for learning Mandarin Chinese as a second language.

6. Conclusions

In the present study, the syllable durations of Mandarin Chinese tones in running speech were compared between native Chinese speakers and Japanese L2 learners of beginner and intermediate proficiency. The patterns of variations of syllable duration by tone reveal the effect of tone type on syllable durations, and the different patterns shown by Japanese L2 learners reveal the problem of generating tones naturally according to Mandarin prosody structures for L2 learners.

The results give us new insight into the issue of tone generation by L2 learners in running speech. Even without separating the syllable positions in utterances, the differences could be clearly observed by looking at the syllable durations for Japanese L2 learners. It is reasonable to assume that the tone generation problems for L2 learners would be clearer if combining the syllable durations with the F0 contours and considering the effects of intonation. These will be the topics of our future research.

7. Acknowledgements

The first author would like to thank Prof. SUNAOKA in Waseda University for her support of speech corpus collection and Nicolas Loerbroks for revising the English.

8. References

[1] J. Wang, "The neutral tone in trisyllabic sequences in chinese dialects," *In TAL-2004*, pp. 201–202, 2004.

[2] A. Jongman, Y. Wang, C. B. Moore, and J. A. Sereno, *Perception and production of Mandarin Chinese tones*. Cambridge University press, 2006.

[3] D. Deterding and S. Xu, "Acoustic investigation of neutral tone in brunei mandarin," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS-18)*, vol. 10, 2015, p. 14.

[4] G. Peng, "Temporal and tonal aspects of chinese syllables: A corpus-based comparative study of mandarin and cantonese," *Journal of Chinese Linguistics*, vol. 34, no. 1, pp. 134–154, 2006.

[5] M. Yip, *Tone*. Cambridge University Press, 2002.

[6] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[7] Y. Xu and M. Wang, "Organizing syllables into groupsevidence from f0 and duration patterns in mandarin," *Journal of Phonetics*, vol. 37, no. 4, pp. 502–520, 2009.

[8] J. Zhang, X. Wang, Y. Sun, M. Nishida, T. Zou, and S. Yamamoto, "Improve japanese c2l learners' capability to distinguish chinese tone 2 and tone 3 through perceptual training," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–6.

[9] T. Zou, J. Zhang, and W. Cao, "A comparison study on f0 distribution of tone 2 and tone 3 in mandarin disyllables by native speakers and japanese learners," in *The 6th International Conference on Speech Prosody, Shanghai China*, 2012.

[10] Y. Xu, "Prosodyproa tool for large-scale systematic prosody analysis," *In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, pp. 7–10, 2013.

[11] C. Cheng, J.-Y. Chen, and M. Gubian, "Are mandarin sandhi tone 3 and tone 2 the same or different? the results of functional data analysis," *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development*, p. 296, 2013.

[12] S. Politzer-Ahles and J. Zhang, "The role of tone sandhi in speech production: Evidence for phonological parsing," *Proc. Tonal Aspects of Languages (TAL-03)*, 2012.