A corpus study of the stress pattern of disyllabic words in Taiwanese Mandarin

1. Introduction

There has been a long debate on whether Mandarin Chinese has word-level stress. Some studies have shown that word-level prominence exists in Mandarin (Duanmu, 2000; Qu, 2013; Feng, 2016), yet it is unclear how stress is phonetically realized in Mandarin. As pitch in tone languages encodes lexical contrasts, duration and intensity are more likely to be the main acoustic cues to stress in Mandarin. These two phonetic cues are often considered to be correlated with the two types of feet respectively: iamb (weak-strong) and trochee (strong-weak). Particularly, intensity difference usually occurs in trochees with the initial syllable being louder, and duration difference occurs in iambs with the final syllable being longer (Hayes, 1995; Hay & Diehl, 2007). Duration has been observed to correlate with Mandarin word-level stress in some perceptual studies. Particularly, duration serves as a cue to grouping syllables into feet (Xu & Wang, 2009), as well as a cue to prominence perception in nonce words by Mandarin speakers (Wagner, Zurita & Zhang, 2021). However, other studies show no evidence for the correlations between duration and word-level stress (Lai, Sui & Yuan, 2010). In addition, some perceptual evidence has shown that the type of tone carried by the syllable has an effect on the perception of stress. Specifically, the final syllable that carries the neutral tone (T5) or the falling-rising tone (T3) is usually perceived as unstressed (Wang et al., 2001). It is thus difficult to pin down what foot type Mandarin has given the mixed findings in the current literature.

Nevertheless, previous studies have exhibit certain limitations. The methodology being used are predominantly through experimental settings, either asking Mandarin speakers to read a list of words (Xu & Wang, 2009), or manipulating the acoustic cues of the stimuli and asking Mandarin listeners to judge which syllable is stressed or more prominent (Wang et. al., 2001; Wagner, Zurita & Zhang, 2021). Very few studies have used a speech corpus that contains a large amount of spontaneous speech. Lai, Sui & Yuan (2010) used 1997 Mandarin Broadcast News (HUB4-NE) Corpus, yet the speech from broadcasters may not well represent speech in a naturalistic setting. Moreover, although the types of tones have been observed to have an effect on stress in disyllabic words, it has not been exhaustively examined how different tonal combinations that carried by the two syllables affect foot construction within the domain of a Phonological Word (see the Prosodic hierarchy in Selkirk, 1986). Finally, the role of intensity in Mandarin word-level stress has been relatively under-studied compared to duration, and some have claimed that intensity has little effects on word-level prominence (Shen, 1993).

The current study uses a speech corpus of Taiwanese Mandarin, aiming to examine syllable duration and intensity differences in the production of disyllabic words. The study also compares the duration and intensity among syllables that carry different tone types. Finally, it explores the effects of tonal combinations carried by the two syllables on their duration and intensity differences.

2. Methodology

2.1 Corpus

The data analyzed in this study come from part of the cross-linguistics corpus constructed by the Origins of Patterns in Speech Lab (OoPS-Lab). The speech samples include both read speech and spontaneous speech, collected remotely from 20 Taiwanese Mandarin speakers (10 females and 10 males), aged from 25 to 61. The speech data were transcribed at word-level (with lexical tonal information) and were forced aligned at the phonemic level using the Montreal Forced Aligner (McAuliffe et al., 2017).

2.2 Data processing

The corpus was first imported into PolyglotDB (McAuliffe et al., 2017), and was enriched with a number of variables, including utterances, which were marked by pauses at least 150ms long. The variable syllable was encoded, for those who contain vowels as their nuclei. A property was then encoded for syllables showing their positions in words ('initial' or 'final'). Information about counts (e.g., syllable count per word) and speakers' demographic information were also encoded. In addition, acoustic measures, in particular, intensity tracks, were also encoded.

I queried all syllable tokens that were in disyllabic words, and filtered out words that were in utterance-final positions. Then I exported the query results (27608 syllable tokens) into a CSV file, with the ID of each syllable token, the label of each token (which syllable it was), the label of the word where the syllable was in, its position in the word (initial or final), the time stamp where it started and ended, the duration of the token, the average intensity of the token based on the measures in the intensity track, and the tone that the syllable carried.

A number of filtering was conducted in RStudio (R Core Team, 2019) based on the data exported from PolyglotDB. First of all, syllables with overly long or short duration were removed. The distribution of syllable duration was plotted on a log scale in histogram using ggplot2 function in R. After careful inspection of the distribution, a lower limit of 0.06s and an upper limit of 0.8s were set to remove the outliers. Since duration was roughly normally distributed in log space, any syllable with a log duration more than three standard deviations away from the mean was removed. Similarly, syllables with overly high or low intensity were also excluded. A lower limit of 35 dB and an upper limit of 86 dB were determined to remove outliers by inspecting the distribution of intensity on a log scale. Any token with a log intensity more than three SDs away from the mean was also removed. The duration and intensity distribution for each speaker was plotted and inspected, and no filtering was conducted within speakers. The final data comprises 26935 syllable tokens from 13789 tokens of word.

3. Results

3.1 The general pattern

In order to examine the overall word-level prominence pattern in Taiwanese Mandarin, duration and intensity of each syllable were plotted using ggplot2 package in R by its position in a disyllable word (initial or final). As shown in Figure 1, the duration of initial and final syllables does not have a significant difference, with the mean of the final syllables (0.210s) slightly longer than that of the initial syllables (0.195s). Figure 2 shows the intensity of initial and final syllables. The mean intensity of final syllables (70.6 dB) is slightly higher than that of the initial syllables (68.9 dB). The overall slightly higher intensity and longer duration of the final syllables suggest that Taiwanese Mandarin tends to have a word-final prominence, yet both of the cues are rather subtle. Recall that trochees usually have higher intensity initially, whereas iambic feet usually have longer duration finally (Hayes, 1995; Hay & Diehl, 2007). Taiwanese Mandarin does not have a clear foot pattern of either trochee or iamb, given that the intensity and duration differences are quite small. Its word-final prominence pattern seems to suggest that it may be a bit iambic-like, yet it is uncommon to see intensity difference as a slightly more robust cue compared to duration for iambic foot.

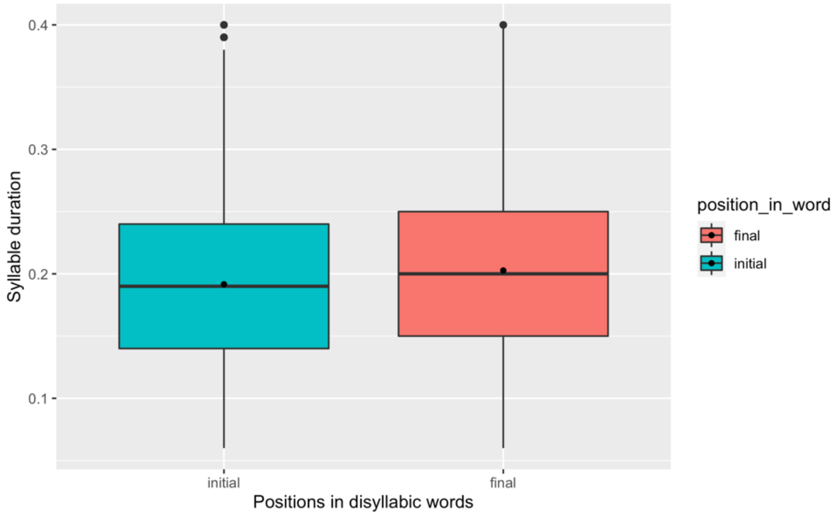

Figure.1. Duration of initial and final syllables

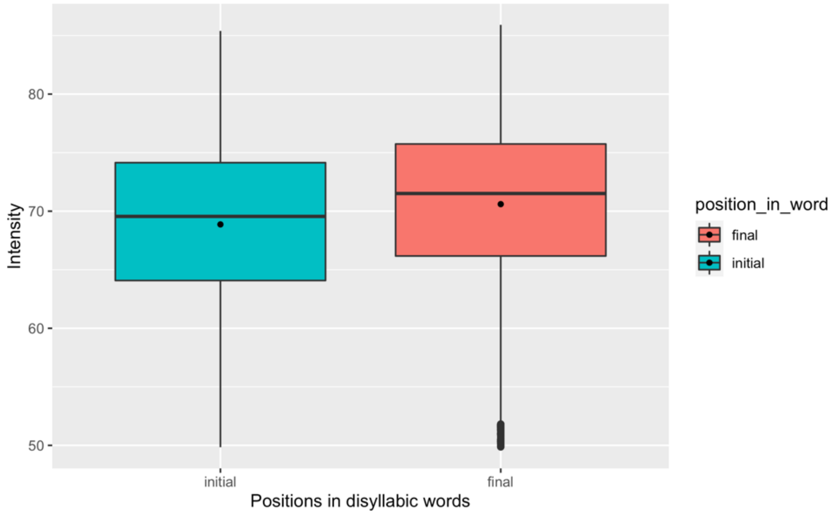

Figure 2. Intensity of initial and final syllables

3.2 The effects of tone types

Previous literature has observed that syllables carrying different tones are perceived to have different levels of prominence by Mandarin listeners (Wang et al., 2001). It is thus important to examine how the acoustic cues of duration and intensity may be realized differently in the production of syllables that carry different tones. Figure 3 shows the duration of syllables with different tone types (T1~T5). Syllables with T5, the neutral tone, are produced considerably shorter than those with the other tones. Syllables with T1, the high level tone, are produced slightly longer than the other contour tones. Interestingly, syllables with T3, the falling-rising tone are produced slightly shorter than the other tones (except the neutral tone), despite its relatively complex contour underlyingly. The third tone sandhi, which refers to a process where the first T3 becomes T2 (the rising tone) when two T3s occur in a row, may have an effect on the realization of T3. Some of the syllables encoded as T3 may be actually produced as T2 instead. It is also possible that T3 is reduced to a low tone by some speakers (Zhang & Lai, 2010), potentially resulting a shortened duration of the syllable that it associates with. These observations suggest that syllables with T5 and T3 are relatively less prominent with respect to their duration cues, and syllables with T1 are the most prominent. This provides production evidence for the previous study that T5 and T3 are usually perceived as unstressed by Mandarin listeners (Wang et al., 2001).

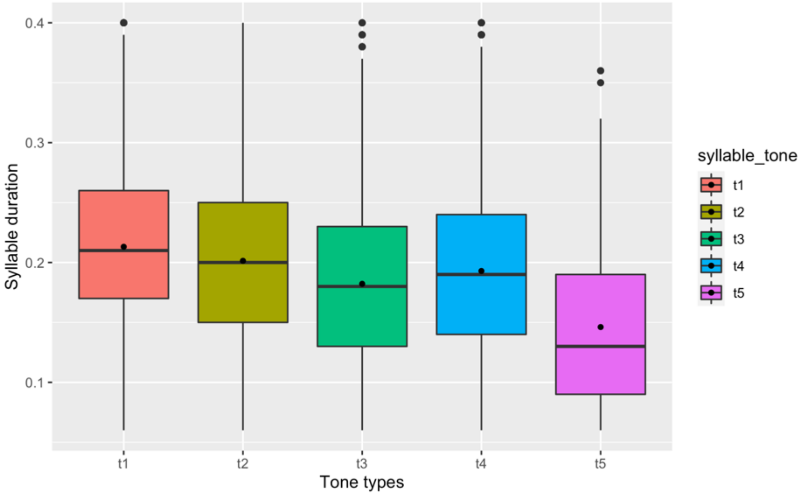

Figure 3. Syllable duration by tone types

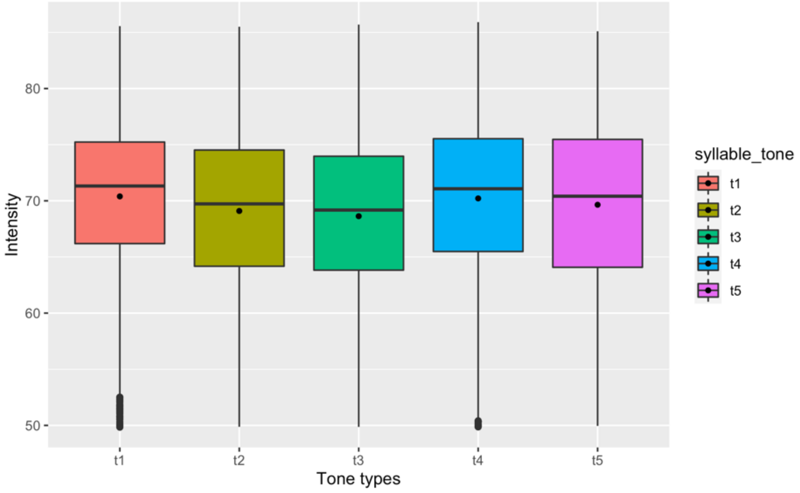

Figure 4. Intensity by tone types

The intensity of syllables carrying different types of tones are plotted in Figure 4. Unlike duration, the type of tones does not have a robust effect on the intensity of the syllable which the tone is carried by. Notably, syllables with T5, the neutral tone, are produced with intensity as high as the other tones. In addition, syllables with T1 and T4 are produced with slightly higher intensity than T2 and T3.

3.3 The effects of tone groups

Given that the production of syllables with different types of tones are cued by intensity and duration in some cases (e.g., syllables with T5 are relatively short in general), it is thus necessary to examine word-level prominence in Taiwanese Mandarin by tonal combinations carried by the words (e.g., T1T2). It is possible to see variations in prominence cued by duration or intensity in words with different tonal combinations.

A variable "tone group" was encoded at word-level based on the tones carried by the two syllables contained in the word. Disyllabic words were then grouped by their tonal combinations (20 different combinations in total, as the neutral tone cannot occur on the first syllable in a word). Duration of the initial and final syllables are measured for each tone group, as plotted in Figure 5. We can see that duration difference between the two syllables vary across different tone groups. Specifically, words ending with T5 tend to have shorter duration on the final syllable, except for T3T5 group. However, the data for T3T5 tone group may be a bit unreliable because of its relatively small number of lexical items (only 6 different lexicons carrying T3T5 in this corpus). Moreover, words ending with T1 tend to have longer duration on the final syllable than the initial regardless of the tone carried by the initial syllable, suggesting an iambic foot pattern. It is hard to conclude any prominence pattern for other tone groups (i.e., words ending with T2, T3 or T4) at this stage, as the duration difference between the two syllables are not very significant.

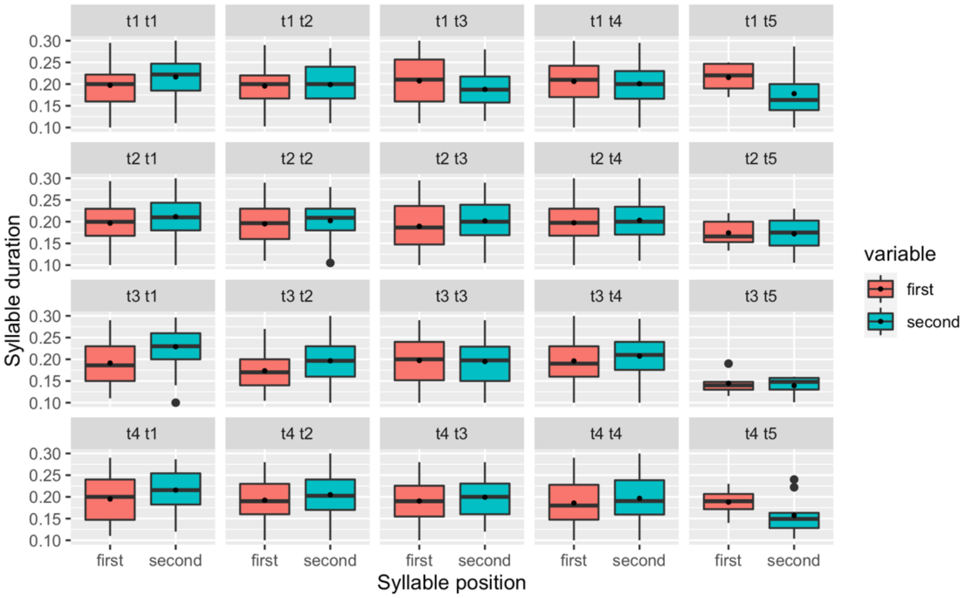

Figure 5. Duration of initial and final syllables by tone groups

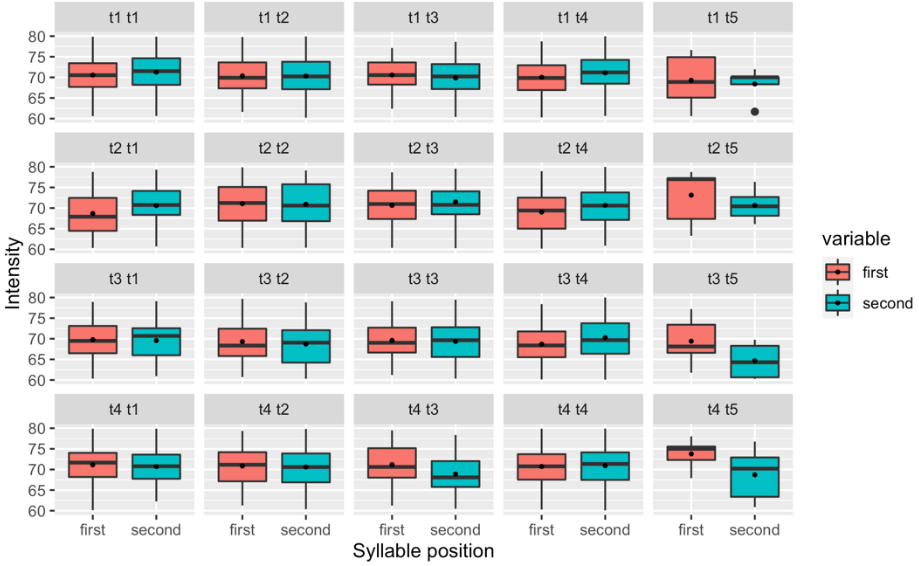

Figure 6. Intensity of initial and final syllables by tone groups

The intensity of the two syllables in disyllabic words is also plotted by different tonal combinations that the words carry, as shown in Figure 6. Words ending in T5 tend to have higher intensity on the first syllable than the second syllable, suggesting a trochaic foot pattern. Words with T4T3 tonal combination also stand out with a trochaic pattern (higher intensity on the initial syllable). There are also several tone groups exhibiting higher intensity on the final syllable, such as T2T1 and T3T4. It seems that there is no systematic pattern across tone groups with respect to intensity difference, except for words ending in T5. Instead, different foot pattern emerges for individual tone group, which may require further careful examination.

4. Discussion

This study uses read and spontaneous speech from a corpus to examine word-level prominence pattern in Taiwanese Mandarin. Although there does not seem to exist a clear iambic or trochaic pattern overall in disyllabic words, duration and intensity, which are potentially cues to stress, are observed to correlate with tones. Specifically, syllables with T5, the neutral tone, are usually cued by shorter duration and lower intensity compared to its preceding syllable in the word, potentially leading to a trochaic foot pattern. Moreover, words ending with T1 exhibit an iambic foot pattern, considering the longer duration in the final syllable carrying T1 compared to the initial syllable. Therefore, fine-grained examination is needed to further analyze how exactly stress pattern correlates with tones in Taiwanese Mandarin. Qu (2013) proposes a weight distinction among the tones in Mandarin based on evidence from phonological processes, namely, T0 syllables are weightless, T3 syllables are light, and T1, T2 and T4 syllables are heavy. This is consistent with the duration difference observed among the five tones as shown in section 3.2.

Moreover, it is difficult to generalize word-level stress patterns in Mandarin without considering the morphological structure of words, given that most of disyllabic words in Mandarin are compound words. The position of stress is likely to correlate with the grammatical relations of the two components. For instance, Feng (2016) provided minimal pairs for Mandarin compound words with the same pronunciation yet different meanings (and different orthography at least for one word), and he found that the positions of stress in these pairs are different based on the judgements from Beijing Mandarin speakers. In the pair *dao.jia* 'Taoism' (adjective-noun) and *dao.jia* 'arrive-home' (verb-noun), the stress is on the first syllable (the adjective) for *dao.jia* 'Taoism', and is on the final syllable (the noun) in *dao.jia* 'arrive-home'. He argued that different stress patterns exist in Beijing Mandarin, which tend to relate to the internal morphological structure of the word, and that stress is specified for certain lexical items.

Compared to using a word list reading method to elicit data (Xu & Wang, 2009), the corpus used in this study comprises a larger scale of speech data with two distinct speech styles (read and spontaneous speech). The present data are also more likely to represent natural speech in real world compared to words that are elicited individually. However, the corpus data may have some limitations given the current question under investigation. Particularly, the word-level stress pattern being examined in this study may be confounded by utterance-level prosodic pattern such as focus, as the words are not elicited in isolation. Word-level stress can be influenced by utterance-level focus; thus, the patterns being observed in this study should take into consideration the potential factor of utterance-level prosody.

Future research could look into the stress patterns of individual tone groups and to see if the correlation between tone and stress will lead to any phonological implications. It is also interesting to take a closer look at how the internal morphological structure of words will influence their stress pattern, and to test if Feng (2016)'s hypothesis for Beijing Mandarin can be applied to other varieties of Mandarin Chinese, such as Taiwanese Mandarin.

References

Duanmu, S. (2000). Stress in Chinese. In *Chinese phonology in generative grammar* (pp. 117-138). Brill.

Feng, S-L. 2016. Beijinghua shi yige zhongyin yuyan [Beijing Mandarin is a language with stress]. *Yuyan kexue [Language Science]*, 15(5), pp.449-473.

Hayes, B. 1995. Metrical stress theory: principles and case studies. Chicago: University of Chicago Press.

Hay J. F. & Diehl R. L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics*, 69(1), pp. 113–122.

Lai, C., Sui, Y-Y. & J-H. Yuan. 2010. A corpus study of the prosody of polysyllabic words in Mandarin Chinese. Proceedings of Speech Prosody 2010. 100457. 1-4.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, pp. 498-502.

McAuliffe, M., Stengel-Eskin, E., Socolof, M., and Sonderegger, M. (2017). Polyglot and speech corpus tools: A system for representing, integrating, and querying speech corpora. In *Interspeech*, pp.3887–3891.

Qu, C. (2013). Representation and acquisition of the tonal system of Mandarin Chinese. Doctorial dissertation. McGill University.

Selkirk, E-O. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, 3. pp.371-405.

Shen, X-N.1993. Relative duration as a perceptual cue to stress in Mandarin. Language and Speech 36 (4). 415-433.

Wang, Y-J, Chu, M, He, L & Y-Q, Feng. 2001. Yuju zhong shuangyinjie yuluci zhongyi ganzhi de chubu yanjiu. [A preliminary study on stress perception of disyllabic prosodic words in utterances]. In Xinshiji de xiandai yuyinxue-diwujie quanguo xiandai yuyinxue xueshu huiyi. Beijing: Qinghua University Press. 166-170.

Wagner, M., Zurita, Iturralde, A. & Zhang, S. (2021). Parsing speech for grouping and prominence, and the typology of rhythm. *Interspeech*.

Xu, Y & M-L. Wang. (2009). Organizing syllables into groups — evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics*, 37, pp.502-520.

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(1), pp.153-201.