A Corpus Study of the Prosody of Polysyllabic Words in Mandarin Chinese

Catherine Lai, Yanyan Sui, Jiahong Yuan

Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA

{laic,yanyan,jiahong}@ling.upenn.edu,

Abstract

This paper presents a corpus study of polysyllabic words in Standard Mandarin Chinese. In particular, this study investigates their prosodic features with respect to the notions of prosodic strength and stress. We find a robust strong-weak alternation with respect to F_0, but different patterns for duration. In disyllabic words the first syllable tends be slightly longer than the second. However, for three and four syllable words the last syllable is the longest, followed by the first. These patterns suggest that F_0 is a reliable phonetic indicator of metrical structure in Mandarin Chinese, rather than duration.

Index Terms: Mandarin Chinese, corpus, tone, prosodic strength, stress, duration.

1. Introduction

This paper presents a study of the prosodic properties of polysyllabic words in Mandarin Chinese. Polysyllabic words form the basic components of speech for which we can look for prosodic regularities and, more specifically, probe notions like rhythm and stress. Although Mandarin lacks lexical stress, various works have argued that words in Chinese still have specific stress patterns beyond the neutral/non-neutral tone distinction. However, there is no consensus about the nature of stress in Mandarin.

Chao [1] claims that the last syllable of a polysyllabic word with non-neutral tone has the 'loudest stress' and the first next. Similarly, Lin et al. [2] found the last syllable duration to be longer than the first in 103 isolated disyllabic words and 154 trisyllabic words with non-neutral tones. Also, the pitch contour of the last syllable approximated the pitch contour in isolation, leading them to conclude that the last syllable is the stressed position. However, the longer observed duration of the last syllable may be due to pre-pause lengthening since isolated words were used in this study. In fact, Wang and Wang [3] found the first syllable of disyllabic words to be longer than the second one in a study of 83 disyllabic words in frame sentences.

The perception of stress seems related to a more general intuition that some elements of speech are "stronger" than others. Recent intonation models of Mandarin tone, such as Stem-ML [4] and PENTA [5], assume that syllables in Mandarin are associated with different strengths. That is, on the one hand, F_0 on syllables with high strength should more precisely approximate canonical tone forms and F_0 swings will tend to be larger. On the other hand, weak syllables can be realized quite differently from their templates. Xu [5] argues that the first and last syllables of a words with more than three syllables are strong. However, Kochanski et al. [4] argue that Mandarin words have alternating strong and weak syllables based on their strength parameter estimation.

From the phonological side, Duanmu [6] proposes initial stress for disyllabic words but in the assignment of compound stress the non-head is stressed. This predicts the second syllable of a Verb-Object (VO) disyllabic compound should be stressed. Wang and Feng [7] propose that the stress in Mandarin is expressed by the different realization of tones at weak and strong positions. They argue that stress falls on the initial syllable of disyllabic words in Beijing Mandarin. This claim is based on native speaker judgements that left strong words are unacceptable if pronounced with the tone pattern of right strong words. However, it is unclear if the unacceptability is due to a real tonal contrast on strong and weak positions, or some other factors, which might have influenced the speakers' judgment when the stress patterns were switched. Nor is it clear how the stress pattern was switched for the perception experiment. Wang and Feng also argue that left stress in Beijing Mandarin is specified in the lexicon, while right stress is predictable from rhythmic, syntactic and semantic considerations. Again, the right strong claim for the VO construction has not been confirmed by phonetic experiment.

The studies above highlight the difficulty in defining and determining the presence of stress in Mandarin. Different acoustic cues, particularly duration, have been claimed to be related to the perception of stress. However, the data used in these studies has been mostly confined to intuitions and small sets of read speech. It is unclear how general the various predictions made by these studies are. In fact, without a clear definition of stress, intuitions about stress could be referring to very different aspects of prosodic structure and rhythm.

In this paper we investigate the prosodic patterns of polysyllabic words in a large speech corpus. Given the various claims above, we would like to know what prosodic and rhythmic regularities exist in words of Mandarin. Our approach is to examine the data without any prior assumptions about the presence of stress. Instead, we aim to show general patterns that arise out of a very large data set and to relate this to previous claims and findings.

2. Data and Methodology

We analyzed speech data from the 1997 Mandarin Broadcast News (HUB4-NE) corpus (LDC98S73, LDC98T24). This corpus consists of recorded news broadcasts from 28 news broadcasters speaking Standard Mandarin. 69% of the words in the corpus were disyllabic, 9% were trisyllabic, and 3% were quadrisyllabic. The transcripts are segmented at the word and utterance levels. Syllable and word boundaries were obtained by forced alignment of the audio to the transcripts using the Penn Phonetics Lab forced aligner [8]. F_0 values for each utterance were extracted using `praat` using the method described in [9]. F_0 measurements were made at 10ms intervals and values were interpolated over voiceless regions. The values were then trimmed and smoothed [10] and then normalized to a semitone scale using the speaker's pitch range.

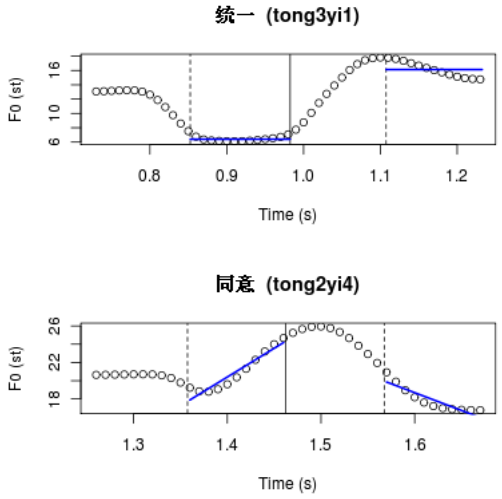

Figure 1: *Tones and F_0 features. The solid vertical line shows the syllable boundary, dashed lines show syllable half-way points. F_0 features were calculated over the second half of each syllable. T1 and T3 are represented by mean F_0 (e.g. tong3yi1, top), T2 and T4 are represented by slopes (e.g. tong2yi4, bottom).*

Mandarin Chinese is a tone language with four lexical tones: high (T1), rising (T2), low (T3) and falling (T4). Thus, in order to view relative differences in F_0 for different syllable positions we need to use measures that account for the fact that Mandarin tones appear to include both dynamic and static targets. For example, we can capture the difference between T1 and T3 as a difference of mean F_0 height, but this will not give a good representation for T2 and T4 rises and falls. We assume that tones are synchronized with the syllable. However, the F_0 from the first part of a syllable represents the transition from the last tone target rather than the target for that syllable [11, 12]. As such, we measure F_0 over the second half of the syllable. T1 and T3 are represented as the mean F_0 over this time period, while T2 and T4 are represented by their linear regression slopes. Figure 1 shows examples of the contours of the four tone types and also demonstrates our F_0 approximations for two disyllabic words taken from our corpus. We assume that stronger T3s and T4s result in lower means and more negative slopes respectively. Hence, we reversed the sign of these measurements so that higher values represent greater strength.

For each syllable, the F_0 measure (mean or slope) was then converted to a z-score based on the mean and standard deviation over all syllables in the corpus bearing that tone. Durations were also converted to z-scores in this way to account for differences in intrinsic tone duration. Z-scores put the variation in the data from different categories onto the same scale. The normalization data excluded T3s preceding another T3 to avoid complications due to third tone sandhi. Some tokens were excluded due to their small number of F_0 points. All in all, this procedure left us with 56378 disyllabic tokens, 4540 trisyllabic tokens and 727 quadrisyllabic tokens. Since we are not focusing on the neutral/non-neutral tone distinction we do not report results for neutral tone syllables in the following.

3. Results

3.1. Word Position

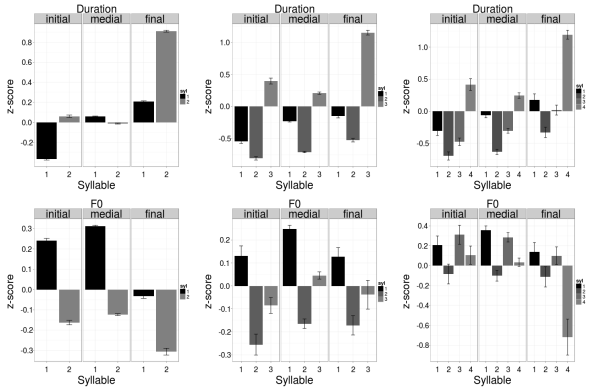

Figure 2: *Duration (top) and F_0 (bottom) means for two, three and four syllable words (left to right) grouped by utterance position.*

As expected, we found significant duration differences between utterance final and non-final words. This lengthening effect is clearly visible in Figure 2 which shows the differences in duration for words in different positions in the utterance. This is congruent with many studies describing lengthening due to prosodic structure boundaries [13, 14]. Utterance final lengthening is clearly separate to inherent prosodic patterns of polysyllabic words. In fact, it is highly likely that this contributes to the perception that the final syllable in a disyllabic word is longer, and hence stressed, when considering isolated words. We see that the utterance initial position also appears to affect the syllable duration. The first syllable of the utterance tends to be shorter than the first syllable of a medial word, for example. To discount these utterance level effects, we will consider only words that appear utterance medially.

3.2. Disyllabic Words

Figure 3 shows differences in F_0 and duration for disyllabic words by syllable position. This shows a general strong-weak pattern for F_0. We can also see a long-short tendency in terms of duration. However, the duration difference appears to be much smaller than that of F_0. We also see this same pattern in Figure 4 which shows syllable differences grouped by tone. This gives us a general picture of how a single tone behaves in different syl-

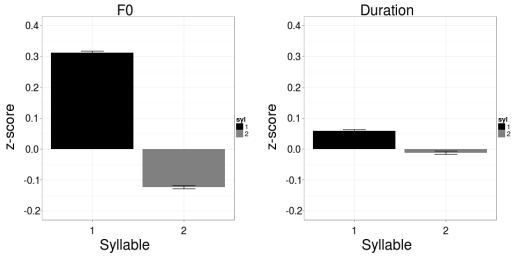

Figure 3: *Disyllabic words. Means and standard errors for F_0 (left) and duration (right) grouped by syllable position.*

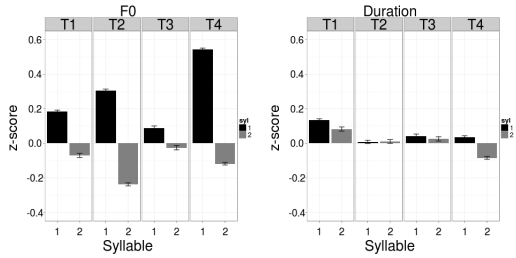

Figure 4: *Syllable differences for disyllabic words grouped by tone: F_0 (left), duration (right).*

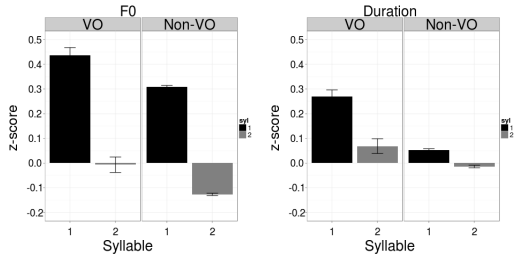

Figure 5: *VO words: F_0 (left) and duration (right)*

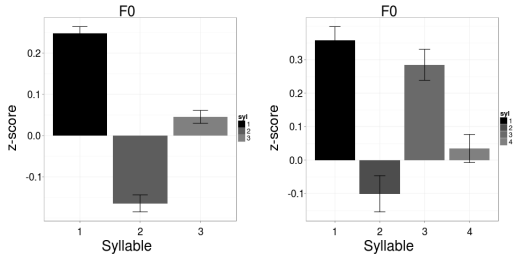

Figure 6: *F_0 means and standard errors for each syllable position for three and four syllable words.*

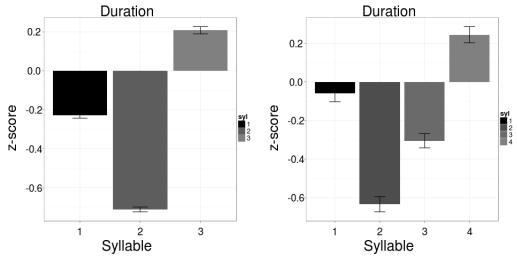

Figure 7: *Duration means and standard errors for each syllable position for three and four syllable words.*

lable positions over the tonal contexts. Again, the strong-weak F_0 pattern emerges for all tones while the duration difference is very small.

F_0 differences are all significant (t-test: $p < 0.001$) for each tone group. It is also clear that each tone has different characteristics. For example, the difference between T1 as a first or second syllable seems much smaller than for T2 or T4. Also, T3 shows a smaller difference between first and second syllable means than T1 (0.8 st versus 2 st for T1). The syllable mean duration z-scores of T2 and T3 are not significantly different (t-test: $p > 0.1$). The second syllable for T1 and T4 does appear significantly shorter than the first ($p < 0.001$). However, the real differences are very small (3 and 7ms for T1 and T4 respectively). So, disyllabic words do not show a strong duration pattern, unlike what we see for F_0.

As noted above, other phonological studies have argued that stress positions depend on the internal structure of the compound. So, although disyllabic words as a group do not show a strong duration pattern, certain types of words might. In particular, they claim that VO compounds should exhibit second syllable stress. The second author, a native speaker of Mandarin, identified the VO words in our data set. The duration and F_0 differences for VO and non-VO words are shown in Figure 5. This indicates that VO words pattern like non-VO words for both F_0 and duration. Specifically, the second syllable is not longer than the first. However, we can still reconcile these results with the intuition that VO words are strong on the second syllable. First, we note that the apparent final stress on VO words derives from the fact that the second syllable is not reduced while this can be the case for other disyllabic words. So the real comparison should be between the second syllables of VO and non-VO words rather than between the syllables within VO words. We can see from the graph that the second syllable duration of VO words is significantly longer than that of non-VO words ($p < 0.001$). That is, the second syllable in VO words is stronger than in non-VO words although the VO disyllabic pattern itself is not necessarily weak-strong.

Thus, in general, the F_0 data supports the view that the first syllable in a disyllabic word is prosodically strengthened. That is, the range of level tones is widened and the magnitude of contour tone slopes is greater. Also, at this point, in utterance medial position we do not find any evidence that disyllabic words exhibit second syllable stress.

3.3. Three and Four Syllable Words

The overall patterns for three and four syllable words are shown in Figures 6 and 7 for F_0 and duration, respectively. These show the same strong-weak alternating pattern for F_0 as was seen for the disyllabic words. However, for duration, the last syllable is the longest followed by the first. That is, the syllables on the word boundaries are longer. We can clearly see that the "strength" in duration does not necessarily mean strength in F_0 or vice-versa.

4. Discussion

Our study has shown a robust strong-weak F_0 alternation in both disyllabic and polysyllabic words in standard Mandarin. In terms of duration, the first syllable of disyllabic words is only slightly longer than the second syllable, but in trisyllabic and quadrisyllabic words, the last syllable is the longest, followed by the first syllable. Clearly, the duration pattern does not show the same alternating pattern that we see for F_0 in three and four syllable words. So, it appears that these two features are not just two aspects of the sort of word level stress discussed in the literature. Different behaviors for F_0 and duration have also been observed in a corpus study of English word stress by Yuan et al. [15], where it was found that the pitch of secondary stress has similar patterns with reduced vowels compared to primary

stress, but with respect to duration, the secondary stress syllable is more similar to primary stress than to the reduced vowels.

The strong-weak F_0 pattern is inline with the alternation of prosodic strength values found in [4] for modelling Mandarin Tone. This alternation also lends support to the trochaic foot structures proposed by Duanmu [6, 16] for Mandarin. In light of the robustness of this alternation in both disyllabic and polysyllabic words, we claim that F_0, as it varies in strength, reveals the underlying metrical structure of Standard Mandarin [17]. As such, F_0, rather than duration, appears to be the phonetic correlate of this sort of prosodic prominence.

Duration appears to mark a different sort of structure. The variation in this feature appears more subject to the effects of prosodic boundaries, as is indicated by the longer duration of the second syllable in disyllabic words at utterance final position. In general, duration and lengthening have been linked to higher level prosodic boundaries [18]. This suggests that the longer duration in the last syllable of three and four syllable words may also be due to the presence of prosodic boundaries. Prosodic words consisting of more than two syllables have longer duration in the last syllable than prosodic words that are disyllabic. Recall that we do not find word final lengthening in utterance medial disyllabic words. At this stage our corpus is not annotated with respect to the higher level prosodic structure. So, although we have excluded utterance level boundary effect, it remains to be seen whether other prosodic or structural effects can account for the final syllable lengthening in three and four syllable words.

5. Conclusion and Future Work

We presented a corpus study of polysyllabic words from a large speech corpus of Mandarin Chinese. We found a robust strong-weak alternation for F_0 for two, three and four syllable words. This differed from the duration pattern for three and four syllable words, where we found the non-medial syllables to be the longest.

This difference between F_0 and duration patterns highlights the difficulty in determining what exactly listeners are picking out when asked to determine 'stress' in a tone language like Mandarin. Looking at our findings in the light of previous work on stress in Mandarin, it appears that the interplay between these two features affects the perception of stress.

Our findings help tease apart the contributions of F_0 and duration in the prosody of Mandarin Chinese. The strong-weak alternation supports the view that metrical structure in Mandarin rhythm is conveyed through F_0 rather than duration. We expect other prosodic structures influence the duration effects that are present. However, further work is needed to show whether this is in fact the case. Given the different duration for disyllabic and three/four syllable words, we expect that further investigations will also shed light on the prosodic structure hierarchy of Mandarin. We would also expect to find different behaviors for F_0 and duration in other tone languages. Further investigations in these directions, and also of conversational speech, are left for future work.

6. References

[1] Y. Chao, *A Grammar of Spoken Chinese*. University of California Press, 1968.

[2] M. Lin, J. Yan, and G. Sun, "Beijinghua liangzizu zhengchang zhongyin de chubu shiyan [Preliminary Experiments on the Normal Stress in Beijing Disyllables]," *Fangyan*, vol. 1, pp. 57–73, 1984.

[3] J. Wang and L. Wang, "Putonghua duoyinjieci yinjie shichang fenbu moshi [The distributional patters of syllable durations in polysyllabic words in Standard Mandarin]," *Zhongguo Yuwen*, vol. 2, pp. 112–116, 1993.

[4] G. Kochanski, C. Shih, and H. Jing, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–646, 2003.

[5] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3-4, pp. 220–251, 2005.

[6] S. Duanmu, *Phonology of Standard Chinese*. Oxford University Press Oxford:, 2000.

[7] Z. Wang and S. Feng, "Shengdiao duibifa yu beijinghua shuangyinzede zhongyin leixing [The stress types of disyllabic words in Beijing Mandarin by means of tonal contrast methods]," *Yuyan Kexue [Linguistic Science]*, vol. 5, no. 1, pp. 3–22, 2006.

[8] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[9] C. De Looze and S. Rauzy, "Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register Variations and Pause Duration," in *Proceedings of Interspeech'09*, 2009, pp. 2919–2922.

[10] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, pp. 55–105, 1999.

[11] C. X. Xu, Y. Xu, and L.-S. Luo, "A pitch target approximation model for F0 contours in Mandarin," in *Proceedings of The 14th International Congress of Phonetic Sciences, San Francisco*, 1999, pp. 2359–2362.

[12] J. Yuan, "Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling," Ph.D. dissertation, Cornell University, 2004.

[13] C. Tseng and Y. Lee, "Speech rate and prosody units: evidence of interaction from Mandarin Chinese," in *Proceedings of Speech Prosody 2004*. ISCA, 2004.

[14] C. Tseng and Z. Su, "What's in the F0 of Mandarin Speech: Tones, Intonation and Beyond," in *ISCSLP'08: The 6th International Symposium on Chinese Spoken Language Processing*, 2008.

[15] J. Yuan, S. Isard, and M. Liberman, "Different Roles of Pitch and Duration in Distinguishing Word Stress in English," in *Proceedings of Interspeech'08*, 2008.

[16] S. Duanmu, "Stress, information, and language typology," *Yuyan Kexue [Linguistic Science]*, vol. 6, no. 5, pp. 3–16, 2007.

[17] M. Liberman and A. Prince, "On stress and linguistic rhythm," *Linguistic inquiry*, vol. 8, no. 2, pp. 249–336, 1977.

[18] Y. Yufang and W. Bei, "Acoustic correlates of hierarchical prosodic boundary in Mandarin," in *Proceedings of Speech Prosody 2002*, 2002.