Comparison of Tonal Co-articulation between Intra- and Inter-word Disyllables in Mandarin

Xiaodong Wang, Wentao Gu**, Keikichi Hirose**, Qinghua Sun*, Nobuaki Minematsu****

* Department of Electronic Engineering,
** Department of Information and Communication Engineering
*** Department of Frontier Informatics
University of Tokyo, Japan
{wxd; wtgu; hirose; qinghua; mine}@gavo.t.u-tokyo.ac.jp

Abstract

Features of tonal co-articulation in Mandarin speech are studied with a focus on how the word boundary location affects the results. Although there are several previous works investigating how the prosodic features of syllables are affected by the surrounding syllables, most of them selected nonsense syllable sequences as speech material without specific consideration on the word boundary. In the present study, however, a comparison is given on the tonal co-articulation between intra-word and inter-word situations. The speech material is designed in that, in each pair of sentences, the target disyllables share exactly the same tonal context but differ in the position regarding to the word boundary: the boundary locating at the initial of the target or locating at the middle. Mean and range of F0 values are adopted as prosodic features of each syllable, and mean F0's differences between the second and the first syllables of the target are calculated and compared for the sentence pairs. Analysis on all of the 16 disyllabic tone combinations shows the effect of word boundary location on the tone co-articulation is different depending on the tone combinations, especially when the target disyllables include a Tone 2 syllable.

1. Introduction

As a typical tone language, Mandarin (i.e., Standard Chinese) has four lexical tones (five if the neutral tone is also counted) to distinguish the meaning of a syllable. The four lexical tones, henceforth denoted as T1, T2, T3 and T4, in their canonical forms can be represented by 55, 35, 21(4) and 51, respectively in a 5-level tone code system [1]. In continuous speech, tonal syllable plays as the smallest distinctive unit, but the F0 patterns in each syllable deviate significantly from their canonical forms in isolated syllables. This modification of tonal patterns in continuous speech plays an important role in transmitting various linguistic information (e.g., syntactic structure) and paralinguistic information (e.g., emphasis, intention, etc.) [2]. The modification is rather complicated, and in the current study we only investigate the effect of co-articulation of neighboring tones.

For example, the F0 values in each syllable change with the neighboring tones. This well-known effect of tonal co-articulation can be divided into two categories. One is that a tone exerts an influence on the F0 contour of the immediately following syllable, known as carryover effect. The other is that a tone starting with a low F0 raises the F0 of the preceding tone, known as anticipatory effect [3]. In [3], these two effects are claimed to be dissymmetrical, whereas in [4] they are claimed to be symmetrical. The seemingly opposite conclusions may be due to the unnatural speech material used in their analyses: nonsense disyllabic sequence /ma ma/ in [3] and tri-syllabic sequence /pa pa pa/ in [4] without considerations of the position of word boundary. When a word boundary locates in the middle of a disyllabic sequence, especially when it is accompanied by a prosodic phrase boundary, the tonal co-articulation pattern of these two syllables will change.

In this study, the differences in tonal co-articulation patterns caused by the different positions of word boundary are investigated through a quantitative analysis on the speech material with natural context, in which two target syllables with different word constitutions are embedded in a pair of sentences for comparison.

The remainder of this paper is organized as follows: Section 2 describes the speech material and analysis method for the comparative study. Section 3 illustrates the results of the analysis. Section 4 gives discussions on the results with a possible explanation on the observation. The paper is concluded with a summary in Section 5.

2. Method

2.1. Speech material

Five pairs of sentences are designed for each of the 16 disyllabic tone combinations. All the sentences begin with a syllable /ta/. In each pair of sentences, two target syllables are embedded as the second and the third syllables. The two syllables compose a word in sentence A, whereas in sentence B they are in different words (*viz*., there is a word boundary between them). Each pair of sentences also shares the same fourth syllable which is immediately following the target disyllable. For instance, a pair of sentences for the target disyllable of T3T3 is given below.

Sentence A: Ta1 | **mai3 hao3** | gong1 ju4.
(He has bought the tools.)
Sentence B: Ta1 | **mai3** | **hao3** gong1 ju4.
(He buys good tools.)

In this example, the target disyllable consists of /mai/ (first target syllable) and /hao/ (second target syllable), and is preceded by /ta/ (pre-target syllable) and followed by /gong/ (post-target syllable). Symbol "|" indicates the location of word boundary. Hence, the sentences A and B share the same first four syllables in the sentence.

The informants include four native speakers of Mandarin (two males: XW, QS; and two females: TX, ZT). Each of the 160 (4x4x5x2) sentences was uttered twice by speaker XW and three times by the other three speakers in a soundproof booth. Hence, there are altogether 1760 utterances. These utterances were recorded at a sampling rate of 22 kHz.

2.2. Analysis method

The speech data are segmented and labeled manually. F0 contours are extracted by using the toolkit Praat [5].

In the first place, the mean, maximum, and minimum of F0 values for each syllable in the targets are calculated. Then, the difference in the mean F0's between the second and the first syllables (henceforth, mean F0 difference) is calculated. Also, F0 range for each syllable is defined as the difference between the maximum and minimum F0 values in the syllable. By taking the average for the entire syllable period, the mean F0 is less affected by various local noises in F0 values and micro-prosody.

The difference in the mean F0 differences between the two target disyllables in a sentence pair is calculated as follows:

SFb-SFa = (Sb-Fb) - (Sa-Fa),

where Sb, Fb respectively indicate the mean F0 values in the second and the first syllables of the target in sentence B, and Sa and Fa respectively indicate those in sentence A, as shown in Fig.1. Mean F0 difference for sentence B disyllable SFb (= Sb-Fb) works as a parameter indicating the influence of inter-word tonal co-articulation, while that for sentence A disyllable works as a parameter indicating the influence of intra-word tonal co-articulation. Therefore "SFb-SFa" corresponds to the difference between tone co-articulations of the inter- and intra-word cases.

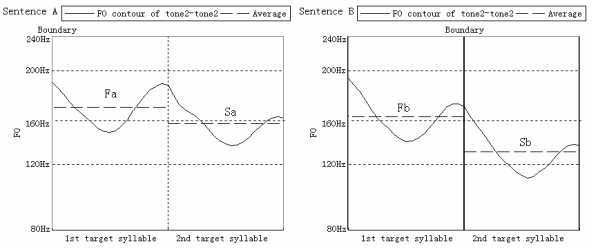

Figure 1. Calculation of SFb-SFa.

Each tone combination of target disyllable has several samples (10 samples for speaker XW and 15 samples for speaker QS, TX and ZT). We calculate mean and standard deviation of "SFb-SFa" for each combination for the further analysis in the next section.

Tukey's HSD (honestly significant difference) Test is one of the multiple comparison tests, which can be used to determine the significant differences between group means. So, with Tukey's HSD Test, the values of "SFb-SFa" for all 16 disyllabic tone combinations are compared together.

In addition, F0 range of target syllables with T2 and T4 are also compared in each pairs of sentences, from another point of view to show tonal co-articulation difference between intra- and inter-word cases.

3. Results

3.1. Comparison of F0 contours

Fig. 2 shows the average F0 contours of target disyllable pairs uttered by speaker QS for each of 16 tone combinations. The first syllable is sorted in row in the order of Tones 1 to 4. The left and right columns respectively indicate the intra- and inter-word cases. Each F0 contour is the average of 15 samples. All syllables are linearly warped to equal lengths before taking the average.

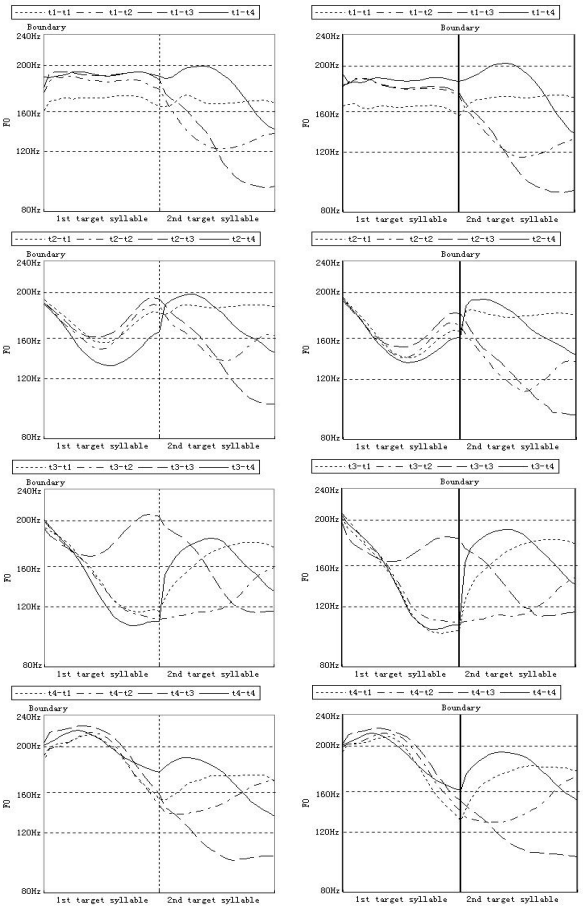

Figure 2. Average F0 contours in the target disyllables for each of the 16 tone combinations.

3.2. Mean and standard deviation of SFb-SFa

Means and standard deviations are listed in Table 1 for each of "SFb-SFa" in the decreasing order of the means. The left, middle and right sub-tables correspond to speakers XW, QS and TX, respectively. Due to space limitations, only the results on speakers XW, QS and TX are presented here, though the discussions in the following sections also apply to speaker ZT. Although not all the tone combinations are exactly in the same order for the three speakers, many of them show more or less consistent orders, as indicated in bold letters. It should be noted that the disyllables with T2 have smaller values than those without T2, and locate at the lower half of the table.

Table 1: Means and standard deviations of SFb-SFa of target disyllable pairs for each tone combination.

Speaker XW			Speaker QS			Speaker TX		
Tone comb.	Mean (Hz)	SD (Hz)	Tone comb.	Mean (Hz)	SD (Hz)	Tone comb.	Mean (Hz)	SD (Hz)
T4T1	23.9	10.4	**T4T4**	28.3	16.8	**T4T4**	10.4	15.3
T4T4	21.1	6.7	**T4T1**	17.5	7.6	**T1T1**	8.4	8.5
T1T1	18.2	5.1	**T1T1**	13.7	5.8	**T4T1**	1.8	8.1
T3T1	17.6	7.4	T1T3	10.1	9.2	T4T3	1.4	26.7
T4T3	13.1	7.9	T3T4	9.8	11.3	T3T1	0.3	19.4
T3T4	12.8	9.0	T4T3	9.7	10.7	T4T2	-1.0	11.7
T1T3	11.9	14.7	T3T1	7.6	11.3	T1T3	-1.3	11.5
T1T4	11.6	8.1	**T1T4**	7.0	7.2	**T1T4**	-2.5	8.9
T3T3	9.5	6.5	T2T1	5.4	6.7	**T3T2**	-6.4	6.8
T4T2	3.7	9.9	**T3T3**	4.4	11.1	T3T4	-7.5	20.7
T2T1	0.3	4.0	T1T2	2.3	8.1	T2T1	-9.6	11.9
T2T4	-1.7	10.5	T4T2	1.3	11.6	**T3T3**	-10.9	22.6
T3T2	-3.5	5.8	**T3T2**	-3.6	7.0	T2T4	-17.0	9.8
T1T2	-5.6	8.3	**T2T3**	-6.8	13.8	**T2T3**	-18.2	15.3
T2T3	-11.2	9.3	T2T4	-7.6	10.9	T1T2	-20.5	27.8
T2T2	-24.6	19.6	**T2T2**	-14.3	7.6	**T2T2**	-39.4	11.1

3.3. Result of Tukey's HSD Test

Since the parameter of SFb-SFa is supposed to indicate the tonal co-articulation difference between intra-word and inter-word cases for each of the 16 disyllabic tone combinations, a comparison of this difference among the 16 combinations can be conducted by Tukey's HSD Test. The results of the tests between T1+T1 and the other 15 tone combinations are shown in Table 2. Those tone combinations showing significantly different SFb-SFa from T1+T1 are marked in bold letters. Additionally, two combinations "T2T1" (for speaker QS) and "T1T2" (for speaker TX), marked in italic letters, are also different from T1+T1 but without enough significance, which will be discussed later.

Table 2: Results of Tukey's HSD Test between T1+T1 and other tone combinations

Tone comb.		Speaker XW		Speaker QS		Speaker TX	
(I)	(J)	Mean (I-J)	Sig.	Mean (I-J)	Sig.	Mean (I-J)	Sig.
T1T1	**T1T2**	**23.82**	**0.000**	**11.35**	**0.104**	**28.86**	**0.000**
	T1T3	6.38	0.985	3.57	1.000	9.66	0.965
	T1T4	6.68	0.977	6.65	0.902	10.85	0.878
	T2T1	**17.90**	**0.002**	*8.25*	*0.627*	**18.02**	**0.101**
	T2T2	**42.85**	**0.000**	**27.94**	**0.000**	**47.77**	**0.000**
	T2T3	**29.42**	**0.000**	**20.40**	**0.001**	**26.62**	**0.001**
	T2T4	**19.88**	**0.000**	**21.30**	**0.000**	**25.39**	**0.001**
	T3T1	0.65	1.000	6.07	0.967	8.08	0.991
	T3T2	**21.68**	**0.000**	**17.21**	**0.000**	*14.77*	*0.508*
	T3T3	8.79	0.796	9.27	0.479	19.29	0.061
	T3T4	5.45	0.997	3.87	1.000	15.92	0.325
	T4T1	-5.67	0.996	-3.88	1.000	6.59	0.999
	T4T2	**14.49**	**0.051**	**12.41**	**0.042**	9.41	0.961
	T4T3	5.10	0.999	4.00	0.999	6.95	1.000
	T4T4	-2.89	1.000	-14.59	0.006	-1.96	1.000

Here "T1T1" indicates disyllabic tone combination of T1+T1, denoted as (I), while the other 15 combinations are denoted as (J). "Sig." indicates the *p*-value of statistical significance for the difference between the means of (I) and (J). When the *p*-value is less than 0.1, the difference is supposed to be significant.

3.4. F0 range

The F0 range of target syllables with T2 and T4 are calculated and analyzed. One of the common results among three speakers is given in Table 3; the F0 range of first target syllable with T2 has a difference between two comparing sentences. Table 3 shows that, when immediately behind word boundary (Sentence A), the first target syllable with T2 probably has wider F0 range than the case it locating before word boundary (Sentence B). The parts shown in bold letters mean that F0 range difference is significant, with significance level as 0.1.

Table 3: F0 range differences for the first target syllable with T2 in various tone combinations.

Tone comb.	Speaker XW		Speaker QS		Speaker TX	
	Mean (A-B)	Sig.	Mean (A-B)	Sig.	Mean (A-B)	Sig.
T2T1	6.58	0.233	**6.38**	**0.083**	**9.00**	**0.025**
T2T2	**9.26**	**0.036**	**10.83**	**0.000**	**7.43**	**0.096**
T2T3	**12.76**	**0.002**	**9.93**	**0.001**	**12.95**	**0.000**
T2T4	7.00	0.182	**6.89**	**0.050**	-3.13	0.823

4. Discussion

4.1. SFb-SFa of T1+T1 and phrasal intonation

Tonal co-articulation and phrasal intonation are two major causes of the variations in tonal patterns of each syllable in continuous speech [6]. F0 pattern of T1 is flat, so mean F0 can reflect its feature well. If it is affected by tonal co-articulation, part or its entire F0 contour will be affected and mean F0 of T1 will also change. For the tone combination T1+T1, two target syllables are both behind word boundary in sentence A, and their mean F0 values are nearly the same, shown in Table 4, which seems to indicate tonal co-articulation of T1+T1 does not result in essential changes in F0 pattern for intra-word case. In addition, both the mean F0 of target syllables are higher than that of pre-target syllable with T1. The low mean F0 of the pre-target may lower the mean F0 of first target syllable due to co-articulation. However, it seems this is not the case, because mean F0's of both the first and the second target syllables are almost the same. We may be able to conclude that at least in the combination T1+T1, tonal co-articulation does not result in essential changes in F0 pattern when two syllables are located at different sides of word boundary.

Therefore, the mean F0 difference between two target syllables with T1 at different sides of word boundary could reflect the value of new phrasal intonation, occurring at the word boundary, which also becomes phrasal boundary here. SFb-SFa of T1+T1 combination just indicates this intonation feature.

Table 4: Mean F0 of the pre-target and the two target syllables for T1+T1 in the sentence A.

Mean F0 of syllable (Hz)	Pre-target	1st target	2nd target
Speaker XW	194.1	**210.1**	**207.8**
Speaker QS	150.8	**161.7**	**158.3**
Speaker TX	271.3	**288.5**	**289.0**

4.2. Difference of SFb-SFa between T1+T1 and tone combinations including T2

Based on the above discussions, the effect of phrasal intonation could be reflected by SFb-SFa of T1+T1, while the other main effect for tonal changes comes from tonal co-articulation. So, if the phrasal intonation effect could be removed, tonal co-articulation will be shown up. This is implemented by finding significant difference of SFb-SFa between T1+T1 and the other 15 disyllabic tone combinations, by referring to Table 2.

For speaker XW, seven disyllabic tone combinations show significantly different results of SFb-SFa from T1+T1. These seven combinations cover all the disyllabic tone combinations which contain T2, i.e. T1+T2, T2+T2, T3+T2, T4+T2 and T2+T1, T2+T3, T2+T4. In the results of speakers QS and TX, most of tone combinations containing T2 also have significant difference from T1+T1.

Therefore, an interesting phenomenon is that nearly all the disyllabic tone combinations including T2 show significant difference in tonal co-articulation between intra-word and inter-word cases, corresponding to two locations of word boundary. The possible explanation will be given as follows.

In Fig. 1 for T2+T2, phrasal boundary of sentence B is at the middle of target syllables and that of sentence A is at the initial of target syllables. So, F0 contour of second syllables in sentences A and B will both be similarly affected to rise by the phrasal intonation. However, the mean F0 of second syllable (Sb) in sentence B is smaller than that (Sa) in sentence A. It seems that phrasal boundary has a depressed effect on the mean F0 of syllable with T2. This depressed effect also occurs for T1+T2, T3+T2 and T4+T2, which could be found in Fig. 2 Therefore, values of SFb-SFa of these combinations come smaller than that of T1+T1.

When the first target syllable is T2, sentence A's phrasal boundary locates immediately before it with the depressed effect. While, with the phrasal intonation's effect for rising on first syllable only in sentence A, the mean F0 value of the first target syllable (Fb) in sentence B is not certain to be smaller than that of the first target syllable (Fa) in sentence A. However, values of SFb-SFa of these tone combinations are still smaller than that of T1+T1.

When both first and second syllables are T2, i.e. T2+T2, with the supposed phrasal boundary 's depressed effect on the first syllable of sentence A and the second syllable of sentence B, the values of SFb-SFa should be decreased much more by the double depressed effect.

Furthermore, the result shows that the value of SFb-SFa for T2+T2 is the smallest in all the tone combinations, as shown in Table 1. Meanwhile, this value is not only significantly smaller than that of T1+T1, just like most other tone combinations including T2, but also significantly smaller than zero, which echoes the double depressed effect by phrasal boundary.

4.3. Relation between tonal co-articulation and syllabic tone pattern

The disyllabic tone combinations in Table 1 show the degree of sum effect of different tonal co-articulation and phrasal intonation effect. Because of text-reading and similar speaking manner for each speaker, phrasal intonation is speculated similar in each sub-table. So, the distribution of these sequences of combinations in Table 1 could reflect some information of tonal co-articulation.

All the seven disyllabic tone combinations including T2 locate in the lower halves in Table 1 for speaker XW and QS, and for speaker TX with one exception, T4T2. This result indicates that T2 has features somewhat different from other tone types. Discussions in Section 4.2 also show the specialty of T2, affected by the phrasal boundary in the inter-word case as compared with the intra-word case.

Moreover, four disyllabic tone combinations consisting of T2 and T3 locate in the lower halves of the two sequences, while another four disyllabic tone combinations consisting of T1 and T4 locate in the upper halves. So, it seems that the degree of tonal co-articulation difference affected by word boundary location has relation to syllabic tone patterns for two groups: T2, T3, and T1, T4.

5. Summary

To compare the tonal co-articulation between intra-word and inter-word cases, speech material is designed in that, in each pair of sentences, the target disyllables share exactly the same tonal context but differ in the position of the word boundary (before or in the middle of the target). Mean F0 and F0 range are adopted as prosodic features of each syllable, and mean F0's differences between the two syllables in the target are calculated and compared within each pair of sentences. Analysis on all the 16 disyllabic tone combinations shows the effect of word boundary location is different depending on syllabic tone patterns.

More quantitative comparative analysis is planned based on the F0 contour generation process model [7].

6. References

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

[2] Ni, J.F.; Kawai, H.; Hirose, K., 2004. Formulating contextual tonal variations in Mandarin. *In Proceedings of ICSLP 2004.* Jeju, Korea, 749-752.

[3] Xu, Y., 1997. Contextual Tonal Variations in Mandarin, *Journal of Phonetics*, 25, 61-83.

[4] Shen, X.S., 1990. Tonal coarticulation in Mandarin, *Journal of Phonetics*, 18, 281-295.

[5] http://www.fon.hum.uva.nl/praat/

[6] Shih, C.L.; Sproat, R., 1996. Issues in Text-to-Speech Conversion for Mandarin. *Computational Linguistics and Chinese Language Processing* 11(1), 37-86.

[7] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E) 5(4)*, 233-242.