Stacking and Unstacking Prosodies: The Production and Perception of Sentence Prosody in a Tonal Language

Cong Zhang

University of Oxford
cong.zhang@ling-phil.ox.ac.uk

ABSTRACT

Teasing apart lexical prosody and sentence prosody has been one of the most difficult tasks in the study of intonational tunes in tonal languages. Are different prosodic manifestations stacked, or are they an integrated whole? With evidence from production and perception data of the intonational yes/no question tune in Tianjin Mandarin at sentence level, this paper proposes that (1) lexical tonal alterations (a.k.a tone sandhi) are lexical-level prosody and do not belong to sentence-level tune; (2) pitch accents induced by information structure are "intra-tune" features, which are such sentence-level prosody features that do not cause sentence type change. Despite being sentence-level prosody features, they are not a part of the tune for intonational yes/no question.

Keywords: intonation; tone; Tianjin Mandarin

1. INTRODUCTION

Pitch modulations are used as cues for both sentence-level and lexical-level prosody in many tonal languages. However, how to tease apart the prosodies at different prosodic levels in production and in perception? Can lexical tones, tone sandhi, pitch accents from the focus of a sentence **ALL** be regarded as components of intonational tunes? If not, what are intonational tunes composed of?

Plenty of studies investigated various aspects of tunes in tonal languages. While most of them describe specific aspects of intonation and argue for the phonological description for each component([1], [2]), few have specifically discussed how to integrate these research results in the prosodic grammar. [3] and [4] have both presented how Mandarin tunes should be analysed or transcribed on the whole, yet a closer look at different sources of the components is still needed. Empirical attempts have been made to study the grammar of different tunes in tonal languages. For instance, [5] investigated the tune of intonational yes/no questions (hereinafter 'IntQ', e.g. '*This is an apple?*' in English) in monosyllabic utterances in Tianjin Mandarin (hereinafter 'TJM', a northern dialect of Mandarin with four lexical tones). They found both a register lift and a floating boundary tone H%[i], which were also used as cues for tune identification. [6] further investigated the chanted call tune of disyllabic calls in TJM, and discovered an actual L% boundary tone at the right boundary of the chanted call tune. However, neither studied whether pitch manifestations other than lexical tones, such as tone sandhi and focus accent, can be directly stacked onto the lexical tones in a sentence tune. The current paper aims to discuss what intonational tunes consist of, by way of studying the IntQ tune in the same tonal language as in [5] and [6], TJM; while the current study investigates longer utterances that contain prosodic units larger than prosodic words. Both the production and the perception results of the current study corroborated with [5] on the register lift and floating boundary H%, and had extra findings on tonal alterations and pitch accents.

2. LEXICAL TONES AND TONAL ALTERATION RULES IN TJM

TJM has four symmetrically distributed lexical tones ([7], [8]): L (T1, low and slightly falling), H (T2, high and slightly rising), LH (T3, low rising), and HL (T4, high falling). Its bitonal combinations have four tonal alteration rules[ii] as in (1), three of which were used in the current study to investigate the relationship between tonal alterations and the tunes.

(1) **a. L→LH/__L**
e.g. *jia*(L) *xiang*(L) → *jia*(LH) *xiang*(L) 'hometown'

b. LH→H/__LH
e.g. *wu*(LH) *dao*(LH) → *wu*(H) *dao*(LH) 'dance'

c. HL→L/__HL
e.g. *da*(HL) *di*(HL) → *da*(L) *di*(HL) 'earth'

d. HL→H/__L
e.g. *bi*(HL) *xu*(L) → *bi*(H) *xu*(L) 'must'

3. PRODUCTION STUDY

3.1. Materials

Two monosyllables ('mao', 'mi') with different lexical tones were used as target words and were used with two different carrier sentences, as shown in (2):

(2) *ta*(L) *shi*(HL)/*xie*(LH) ***mao/ mi*** *zi*(HL)?/.
It/He be /writes mao/ mi character?/.
It is/He writes the character "mao"/ "mi"?/.

ta(L) *shi*(HL)/*xie*(LH) ***mao/ mi*** *zi*(HL)?/.
It/He be /writes mao/ mi character?/.
It is/He writes the character "mao"/ "mi"?/.

The carrier sentences were designed to serve the following three purposes: First, the carrier sentences were used to detach the varying lexical tones from the intonational phrase boundaries, so that the potential pitch accents could be discovered. Second, the carrier sentences created a fixed sentence prominence by way of a syntactic structure that automatically sets the target words as the narrow foci of the sentences; random assignment of foci in broad-focus sentences (see [9]) was thus avoided. The variation of the lexical tones of focused words and the potential associating post-lexical pitch accents then could have a direct interaction. The third purpose of the carrier sentences is to vary the lexical tones of the syllable preceding the prominence to create tonal alteration conditions. This study used two different tones for the syllables preceding the target words – *shi* (HL) and *xie* (LH), to create tone alteration conditions (Rule 1b, c, and d) in order to examine whether tone dissimilation rules have a post-lexical effect on the tunes in TJM.

3.2. Speakers and Procedures

Six native speakers of TJM (3 male and 3 female) were recorded. All speakers were born and raised in the city area of Tianjin and spoke TJM on a daily basis. The reading was recorded using a Rode NT-USB Microphone with Audacity onto a PC, at a sampling rate of 44.1 kHz. The informants were given time to briefly familiarise themselves with the materials before the experiment started. The test materials were presented as Chinese characters without context in a Microsoft PowerPoint presentation. The informants were asked to read form the screen as naturally as they could, by producing a statement when seeing a Chinese full stop "。" at the end of the utterances, and an IntQ when seeing a Chinese question mark "？".

3.3. Results

A mixed-effect model was used to analyse the relationship between various independent variables (duration, mean pitch, F0 range, etc.) and Tune TYPE and TONE. TYPE (Statement, IntQ), TONE (L, H, LH, HL), TYPE: TONE interaction, RHYME ([aʊ], [i]), and GENDER (M, F) were taken as fixed effect factors. SPEAKER (6 different informants) and ITEM (16 tokens) were held as random-effect factors, with intercepts for both SPEAKER and ITEM, as well as by SPEAKER random slopes for the effect of TYPE.

Duration: At the utterance level, the mean duration of the IntQs was longer than that of the statements. The nuclear accents in IntQs were shorter than in statements ($\overline{\text{Dur}}_{[\text{IntQ}]}$ = 274.7ms, $\overline{\text{Dur}}_{[\text{S}]}$ = 296.0ms; TYPE: $\chi^2(1) = 16.10$, $p < 0.001$***), while the post-nuclear accents were longer in the IntQ tune than in the statement tune ($\overline{\text{Dur}}_{[\text{IntQ}]}$ = 320.5ms, $\overline{\text{Dur}}_{[\text{S}]}$ = 281.2ms; TYPE: $\chi^2(1) = 9.16$, $p = 0.002$**). These results indicated that the floating tone at the right boundary added length to the IntQs, while compressing the nuclear parts to highlight the prosodic prominence. This type of final lengthening is boundary-induced (c.f. tune-induced in [6]), which is not likely to be a perceptual cue.

Register: The register of the IntQs was higher than statements as indicated by the mean pitch data (TYPE ($\chi^2(1) = 9.80$, $p = 0.0017$**) and TONE ($\chi^2(3) = 15.01$, $p = 0.0018$**)). Moreover, both maxima and minima were higher in IntQs than in statements (maximum pitch: TONE ($\chi^2(3) = 119.07$, $p < 0.001$***), TYPE ($\chi^2(1) = 7.84$, $p = 0.0051$**); minimum pitch: TONE ($\chi^2(3) = 124.33$, $p < 0.001$***), TYPE ($\chi^2(1) = 12.10$, $p = 0.0005$***), TYPE*TONE ($\chi^2(3) = 72.95$, $p < 0.001$***)). These results indicated that the register was lifted, instead of merely being expanded. The post-nuclear accent had the biggest difference between the mean pitch of the IntQ tune and the statement tune, while the pre-nuclear accent only marginally differed in statements and IntQs. The nuclear accent in the IntQ tune was higher than its statement counterpart, but the difference is not as big as in the post-nuclear accents.

Pitch accents: The alignment data of the pitch maxima and minima of both the IntQ tune and the statement tune were consistent, and did not show any significant overriding pitch accent (maxima alignment – TYPE ($\chi^2(1) = 0.016$, $p = 0.9$, n.s.; minima alignment – TYPE ($\chi^2(1) = 0.68$, $p = 0.41$, n.s. Table 1). Combining the alignment data with the data of mean pitch, pitch maxima and pitch minima, a H* was found to be potentially associated with the prominence.

Table 1. Distances from F0 maxima/minima to the onsets of sentence nuclei

	Tone	Type	Max F0 Dist.	Min F0 Dist.
FALLING	TONE 1 (L)	Q	7.40%	90.68%
		S	3.30%	86.27%
	TONE 4 (HL)	Q	37.92%	85.64%
		S	30.35%	88.93%
RISING	TONE 2 (H)	Q	94.43%	8.58%
		S	92.11%	14.09%
	TONE 3 (LH)	Q	25.42%	47.51%
		S	42.02%	53.06%

The post-nuclear accent went through post-focus compression: the mean F0 range of post-nuclei in IntQs and statements were 1.21ERB and 1.07ERB, while the its statement nuclei counterpart was 2.55ERB on average. However, these differences were results of the focus. The difference in tunes did

not generate any pitch accent difference. Additionally, although tonal dissimilation rules brought a phonetic carry-over effect, they did not create any phonologically meaningful pitch accent or boundary tone.

Edge/ Boundary tone: Consistent with the analysis in [5], the IntQs in the current study also had smaller F0 range than statements. Taking into the duration results of boundary-induced lengthening in IntQ tune, the floating Ḥ% boundary tone analysis was thus also supported.

3.4. Discussion

The results can be summed up as the following six information points:

(a) a boundary-induced final lengthening in IntQ tune;
(b) an overall lifted register in IntQ tune;
(c) a H* pitch accent associated with the sentence prominence;
(d) a compressed post-nuclear accent;
(e) a phonetic carryover effect induced by tonal alteration rules;
(f) and a floating Ḥ% boundary tone in IntQ tune.

Apart from the above six features, lexical tone is also a part of the prosody. Which of these seven prosodic features belong to the tune of IntQ then? The first principle is that the components have to be on the sentence level. Lexical tones are thus excluded. Tonal alteration rules in (e) did not create any phonological change on sentence-level prosody despite being phonological rules themselves on the lexical level. The second principle is to examine the source of the prosodic features. Lengthening could be a sentence-level prosodic features as in the chanted call tune in [6]; however, in the current study, the lengthening was induced by the boundary phenomenon. (a) is therefore also eliminated. (c) and (d) seem to be the most qualified candidates for sentence-level prosodic features; however, they are induced by sentence prominences, which exist both in the statement tune and the IntQ tune. So, such pitch accents caused by information structures can be called 'intra-tune' features instead of 'inter-tune' features. Therefore, only (b) and (f) remain in the IntQ tune expression:

lifted register + floating Ḥ%

4. PERCEPTION EXPERIMENT

Perception data were collected to investigate how the tunes are processed and whether the phonologically meaningful prosodic features are used in identification (i.e. 'unstacking') of the tunes.

4.1 Participants

A total of 28 native Tianjin speakers (13 female; 15 male) participated in this perception experiment. Their age ranged from 20 to 28 years old (mean age = 20.25). None of them had any hearing loss or speech disorders. The participants were not the same participants as the informants in the production study.

4.2 Stimuli and Procedures

The sentences from the production study were used as stimuli for the participants to identify whether they were questions or statements. The experiments were conducted in a quiet room at Tianjin University with the auditory stimuli being played through individual closed-ear headphones (Sennheiser PX200 stereo headphones). The participants were tested in groups of a maximum of five. The participants were asked to decide whether each stimulus was a question or a statement. They made their choices on custom-made individual two-button handsets which were labelled as "陈述。" ('statement' in Chinese, with a Chinese full-stop) above the left button, and "疑问？" ('question' in Chinese, with a Chinese question mark) above the right button. The stickers were switched for left-handed participants to keep the "question" choice on their dominant-hand side. The participants were explicitly instructed to press the buttons with their thumbs and make their responses as accurate and as fast as possible.

4.3 Results

All the accuracy data were analysed with generalised linear mixed models. To analyse the accuracy of the responses, binomial regression models were constructed with the ACCURACY (Correct, Incorrect) as the dependent measure. The fixed factors were TONE (L, H, LH, HL), TYPE (IntQ, Statement), and their interaction TONE * TYPE. PRETONE (LH, HL), SUBJECT, and ITEM were the random factors, with intercepts for all three random factors, as well as by-SUBJECT random slopes for the effect of TYPE. Models including the main effects or the interaction were compared to the same models without main effects or interaction through Likelihood Ratio Tests, where the p-value is derived. When further examination of the interaction between the two main effects was needed, post hoc Tukey HSD tests were conducted. The reaction time data were analysed with linear mixed effect models, with the same main factors and random factors as in the accuracy models.

To simplify the presentation of results, the tones are ranked by their descriptive results: the order of accuracy is descending, while that of the reaction time is ascending. Both orders indicate that on average the

participants made more accurate and faster choices for the ones on the left than the ones on the right. The almost equal sign "≈" is used for symbolising that the two elements connected do not differ significantly from each other. The greater-than and lesser-than signs are used to indicate significant differences.

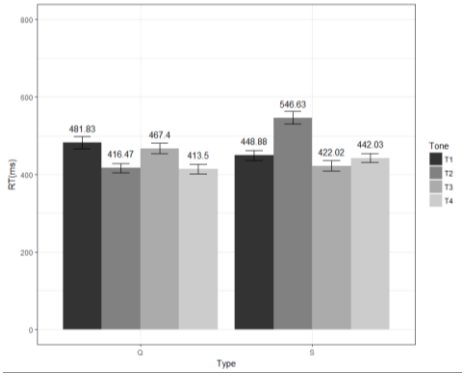
The results of the accuracy exhibited a significant interaction between TONE and TYPE, TYPE*TONE ($\chi^2(3) = 36.21$, $p < 0.001$***). The identification rates were generally high: the accuracy of IntQ tune saw the highest in H and HL, both reaching over 90%; LH and L were lower, but still reached more than 70%. For statements, the accuracy of HL, LH, and L were all more than 90%, but H once again fell to chance level, 53.9%. Combining the descriptive data and the post hoc Tukey HSD test, the results can be demonstrated as in (3) and (4):

(3) IntQ: T2(H) ≈ T4(HL) > T3(LH) ≈ T1(L)
(4) S: T4(HL) ≈ T3(LH) ≈ T1(L) > T2(H)

These accuracy results showed that the starting part of each lexical tone of the target words influences the identification: when the lexical tone of a target word started with a H tone, it was more likely to be identified as an IntQ tune than the low starting tones. Vice versa, in statements, the tones that started low were preferred. The starting tone was the sign of pitch register level. These results indicated that the pitch register was an important cue for the identification of IntQ tune. The reaction time data showed that TYPE, and TONE*TYPE interaction were the significant factors that affected the reaction time of the responses: TYPE ($\chi^2(1) = 0.25$, $p = 0.62$, n.s.), TONE ($\chi^2(3) = 4.76$, $p = 0.19$, *), TYPE*TONE ($\chi^2(3) = 16.62$, $p = 0.0008$***). The trend was consistent with that of the accuracy: HL, H, LH were all shorter than L in IntQs; H was longer than all other tones in statements. It can be expressed as (6) and (7):

(6) IntQ: T4(HL) ≈ T2(H) ≈ T3(LH) < T1(L)
(7) S: T3(LH) ≈ T4(HL) ≈ T1(L) < T2(H)

Figure 1. Reaction time of the identification task



Compared with the reaction times reported in [10] for tune identification of monosyllabic utterances, the reaction times in the current experiment were much shorter (Figure 1). This indicated incremental processing of sentence tunes. Therefore, in this experiment, the boundary information only facilitated the identification when the lexical tones interfered with the identification.

In addition, we also separately examined the effect of tonal alterations rule on the processing of IntQ tune. Two carrier sentences with different tones (LH, HL) preceding the target words were involved in the experiment. Accuracy and reaction time were statistically tested with a mixed-effect model – TONE, TYPE, RULE of tonal dissimilation, and the interaction between TYPE*RULE were the fixed factors. SUBJECT and ITEM were taken as random factors, with intercepts for both SUBJECT and ITEM, as well as by-SUBJECT random slopes for the effect of TYPE. The interaction of TYPE*RULE was the most important factor in the current analysis. The accuracy results showed a significant difference (TYPE*RULE ($\chi^2(3) = 30.97$, $p < 0.001$**) between the tones that went through dissimilation and the tones that did not. In the statement tune, whether there were dissimilation rules or not did not create a huge difference, although the ones that went through tonal dissimilation had slightly lower accuracy by 3.7%. In the IntQ, however, this discrepancy was enlarged to 17.1%. The reaction time data did not differ between the two conditions (TYPE*RULE ($\chi^2(1) = 2.64$, $p = 0.104$, n.s.). These results revealed that tonal dissimilation rules, despite being on the lexical level, placed more cognitive burden on the identification, especially in IntQs which were already more difficult than the statements.

5. SIGNIFICANCE ANF FUTURE STUDIES

The significance of this study is three-fold: firstly, the results support the monosyllabic IntQ analyses in [10]; secondly, it provides a closer examination of other aspects that monosyllabic utterances could not incorporate; lastly, the current study provides insights to the questions in the beginning of this paper: how the intonational tunes are stacked and how listeners unstack them during tune processing? To answer, we propose: (a) Tone sandhis are lexical, so they do not belong to the intonational tunes. Nevertheless, they do affect tune processing as extra cognitive load. (b) Pitch accents induced by information structure are "intra-tune" features, which are such sentence-level prosody features that do not cause sentence type change. They should not be included in the tunes despite being sentence-level prosody features, since intonation tunes should be described contrastively and minimally. An ongoing further investigation is on an "inter-tune" prosody stacking of list intonation and IntQ to analyse more complicated interactions.

5. ACKNOWLEDGEMENTS

I thank the Chinese Scholarship Council for the sponsorship of the study, and my supervisor Professor Aditi Lahiri for her guidance and support. I also thank Dr. Lei Liang at Nankai University and Dr. Hui Feng at Tianjin University for making the recording at their universities possible.

6. REFERENCES

[1] G. Bruce, *Swedish word accents in sentence perspective*, vol. 12. Lund: Gleerup, 1977.

[2] B. A. Connell and D. R. Ladd, "Aspects of pitch realisation in Yoruba," *Phonology*, vol. 7, no. 01, pp. 1–29, 1990.

[3] D. R. Ladd, *Intonational phonology*. Cambridge, MA: Cambridge University Press, 1996.

[4] S. H. Peng, M. K. Chan, C. Y. Tseng, T. Huang, O. J. Lee, and M. E. Beckman, "Towards a Pan-Mandarin system for prosodic transcription," in *Prosodic typology: The phonology of intonation and phrasing*, S.-A. Jun, Ed. Oxford, UK: Oxford University Press, 2005, pp. 230–270.

[5] C. Zhang, "Tones and Tunes in Tianjin Mandarin," in *Tone and Intonation in Europe*, 2016.

[6] C. Zhang, "Chanted Call Tune in Tianjin Mandarin: Disyllabic Calls," in *9th International Conference on Speech Prosody 2018*, 2018, pp. 522–526.

[7] Q. Li, Y. Chen, and Z. Xiong, "Tianjin Mandarin," *J. Int. Phon. Assoc.*, pp. 1–20, 2017.

[8] J. Zhang and J. Liu, "Tone sandhi and tonal coarticulation in Tianjin Chinese," *Phonetica*, vol. 68, no. 3, pp. 161–191, 2011.

[9] Y. Xu, "Effects of tone and focus on the formation and alignment of f 0 contours," *J. Phon.*, vol. 27, no. 1, pp. 55–105, 1999.

[10] C. Zhang, "Tianjin Mandarin Tunes: Production and Perception data," in *Phonetics and Phonology in Europe 2017*, 2017.

[11] L. M. Hyman and M. Tadadjeu, "Floating tones in Mbam-Nkam," *Stud. Bantu Tonology South. Calif. Occas. Pap. Linguist.*, no. 3, pp. 59–111, 1976.

[i] Following the tradition in tonal literature (such as in [11]), a circle under a tone represents the notion of a 'floating' tone.

[ii] 'Tonal alteration rules' are more commonly known as 'tone sandhi rules'; yet, '*sandhi'* means 'assimilation' in Sanskrit, while the rules in TJM are dissimilation rules.