A Tone-based Hierarchical Structure of Chinese Prosody

Ya Li
Zhejiang Ocean University/ No. 1 Haida Road, Zhoushan, Zhejiang 316022
liyawy@zjou.edu.cn

Abstract

In Chinese speech engineering, many projects use a conventional, syllable-based prosodic hierarchy as an underlying framework to process natural or synthesized speech. However, Chinese as a tone language has its own way of expressing prosody, that is, through tonal interaction, especially tone sandhi. By utilizing the capacity of tone as a dual unit of pitch and timing, the present study proposes a tone-based, three-layer-four-level structure for Chinese prosody. The three layers are tone, tone prosody, and intonation, respectively composed of one level of pitch units, two levels of tone prosody units (basic and derived), and one level of intonation units. These four levels of units are used to replace syllable, prosodic word, phonological phrase, and intonational phrase in a conventional hierarchy. Tone prosody units are established based on sizes or types of tone sandhi domains, so when applied to the same clause uttered in Mandarin and Shanghai Wu Chinese, they are timed differently and branched toward different directions at different levels, hence capable of capturing rhythmic and melodic patterns of the two distinctive types of Chinese. Overall, given its theory-friendly design, the proposed structure may be used as a unifying framework in Chinese speech engineering.

1 Introduction

In speech engineering, it is notoriously difficult to improve the naturalness of synthesized speech due to multiple linguistic and non-linguistic factors involved. It is especially true for Chinese speech synthesis, because it is hard to find a unifying prosodic framework optimized for the whole Chinese language family, both standard Mandarin Chinese and other Chinese dialects. At present, most models assume a syllable-based prosodic hierarchy developed for non-tonal languages. Such a hierarchy is conventionally divided into eight layers from bottom to top: mora, syllable, foot, prosodic word, clitic, phonological phrase, intonational phrase, and utterance (Nespor and Vogel, 1986). This hierarchy is built around stress foot, hence readily applicable to stress languages like English. For Chinese, according to Cao (2011), a simplified version of this conventional hierarchy is generally assumed, and it has three basic levels, prosodic word (PW), prosodic phrase (PP), and intonational phrase (IP). However, the applicability of such a simplified hierarchy to Chinese is still questionable, because Chinese has tone rather than stress on its syllables and the status of tone is not reflected in this hierarchy.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Alternatively, Zhang (2017) proposed the prosody-syntax sensitivity model of prosodic hierarchy based on tone sandhi domains. He used examples from Xiamen Min Chinese to show that syntactic information directly participates in the whole process of tone sandhi, from domain construction to sandhi execution. His model acknowledges the role of tone sandhi domains in mediating between syntax and prosody, yet it is still based on the conventional, syllable-based hierarchy. The present study, on the other hand, holds that any Chinese prosodic hierarchy should involve tone directly, as it is the most important feature of Chinese speech. It thus proposes a tone-based hierarchical structure for Chinese prosody.

The present paper first introduces research background (Section 2), followed by an illustration of the proposed structure (Section 3). Then, examples of its application are presented (Section 4) and comparisons are made between the proposed structure and other structures (Section 5). Last, implications of the proposed structure for prosody modeling are discussed (Section 6), followed by a brief conclusion (Section 7).

2 Research background

Similar to Zhang (2017), who claimed that the Chinese prosodic hierarchy must refer to tone sandhi domains, the present study holds that a tone-based prosodic hierarchy must be built around tone sandhi domains. According to Yan (2010), tone sandhi domains are hierarchically ordered and tone sandhi rules take prosodic units as their domain of application: Some take prosodic word while others take phonological phrase. Nonetheless, this is not always the case with all Chinese dialects.

Worse than all, the notions of word and phrase in Chinese are problematic, as they are very different from those in English: a Chinese *zi*, a mono-syllabic meaning unit, is usually a word by itself. A multi-*zi* sequence, when grammatical markers are lacking, can be used as a word, a phrase, or a clause, depending on the pragmatic context. Zhuang (2015) suggested that a Chinese word should be viewed as a unit across multiple prosodic layers since there is no clear boundary between a word and a phrase. Shi (2017) also proposed a dual unit called Chinese *zi*-word to encompass both *zi* and word. Establishing a tone-based prosodic hierarchy for Chinese may help differentiate *zi*, word, and phrase if each of them has a distinctive tone sandhi pattern. Better yet, Li (2023) in her study of tonal timing proposed a dual pitch-timing structure for tone, which allows tone to serve as a basic prosodic unit instead of syllable. Moreover, Xu (2018) studied English stress and Chinese tone as well as their respective relationship with rhythm and concluded that Chinese rhythm should be tone- rather than syllable-based.

Here there are two notions, melody and timing, to be clarified. According to Hirst (2013), melody refers to a global pitch fluctuation pattern shaped by tone, stress, and intonation and measured by scope and speed of change in pitch height. A faster and larger pitch fluctuation leads to a more melodious speech. Timing is associated with rhythm. Rhythm refers to a recurring timing pattern of prosodic units, a regular duration change caused by syllable type, tone, stress, pause, and speech rate and measured by scope of change in segmental duration. It generally falls into three classes: stress-timing, syllable-timing, and mora-timing (Ladefoged and Johnson, 2011). In theory, they respectively have stress foot, syllable, and mora equally timed during speech. In reality, their timing distinctions are not fully supported by experimental data. Nonetheless, different languages do show different tendencies toward one class or another.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

For example, Hirst (2013) found that Mandarin Chinese is more syllable-timed and also more melodious than French, and French in turn more syllable-timed and also more melodious than English. In addition, English declarative sentences show an overall falling melody whereas French ones an overall rising melody. Li (2015) conducted a further study of four Chinese dialects, Northern (including Mandarin), Cantonese, Wu, and Min. Her results show that the four dialects all tend to be syllable-timed and melodious, but to different degrees: Cantonese and Northern dialects are more syllable-timed and also more melodious than Wu and Min. She attributed these differences to their distinctive tone inventories and tonal behaviours.

Last, many experimental studies find that different phonetic cues are used to signal boundaries of different prosodic units (Yin, 2020). In Mandarin, an intonational phrase is cued by pitch height decline towards the end, followed by a long pause; and a phonological phrase by final lengthening. Prosodic words, on the other hand, do not have clear boundary cues, but they can be cued by tone sandhi.

3 The proposed prosodic structure

The underlying assumption of the proposed structure is that the pitch-timing duality of tone is capable of directly or indirectly reflecting Chinese melodic and rhythmic patterns. In other words, tone alone is sufficient to characterize Chinese prosody. Therefore, the syllable layer in the conventional prosodic hierarchy can be replaced by the tone layer. Also, the syllable-based stress foot can be replaced by a tone-based unit, called a tone prosody unit (TPU) in the present study. A tone sandhi domain is a typical TPU. Since tone sandhi is closely related to morpho-syntactic conditions, a sandhi-domain-based TPU helps mitigate the need to distinguish words from phrases. TPUs can be divided into two types and placed on two levels. There is good reason to differentiate them in terms of levels, which will be explained soon. Now the tone-based prosodic structure can be illustrated below:

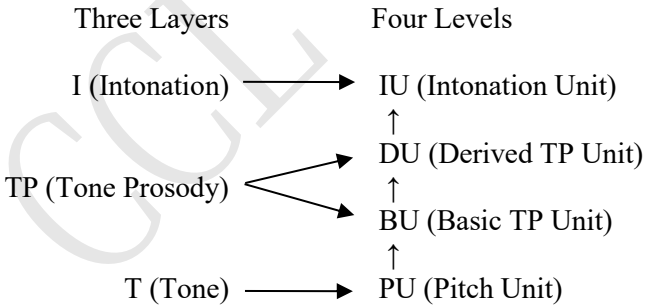


Figure 1. The Tone-based Prosodic Structure of Chinese.

In Figure 1, the tone-based structure has three core layers: T, TP, and I. They are composed of different prosodic units. PU refers to different pitch height components, usually represented by pitch levels, high, mid, and low (H/M/L), or pitch values from 5 to 1 on a five-level scale. For example, the Mandarin fourth tone (T4), a falling tone, is usually represented as HL or 51, indicating that it contains two PUs, H and L. Note that if a BU contains two tones, a tonal interaction process (including tone sandhi) will be evoked to reduce the PU number due to the timing pressure, but the PU in the head position will be retained due to the need to create prosodic prominence (Li, 2023). Take the Mandarin third tone (T3) sandhi for example. If a BU contains a T3.T3 sequence (LLH.LLH), it will undergo the T3 sandhi, turning T3.T3 into a sequence similar

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

to T2.T3 (LH.LLH). In connected speech, the resulting pitch sequence LH.LLH is often reduced to L**H**L, where the central **H** in the head position is retained.[2]

Next, the structure has BU and DU in the same TP layer but on different levels. They are in the same layer because they both can be domains of tone sandhi and share some structural features; for example, they both can be recursively built by adjoining another tone, a process similar to cliticization. Their division, on the other hand, is for the purpose of accommodating tone sandhi domains of different sizes or types. Later, examples will be given to show their differences. Generally, they encompass the conventionally defined prosodic word, clitic, and phonological phrase, but there is no strict one-on-one relationship between BU and prosodic word as well as between DU and phonological phrase. Last, the upper level IU is equivalent to the conventionally defined intonational phrase. Overall, the proposed structure is aimed to capture Chinese melodic and rhythmic patterns through interactions of different units both within the same or across different layers. Now Mandarin examples are presented to show how the four levels of prosodic units are formed:

(a)BU as a tone sandhi domain:	(T3.T3)$_{BU}$ → T2.T3 *shui.guo*, 'fruit'
(b) BU' (Recursive BU):	((T3.T3)$_{BU}$T0)$_{BU'}$ → T2.T3.T0 *shui.guo.de,* 'of fruit'
(c)DU:	[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$ → T2.T3-T2.T3 *shui.guo-pin.zhong,* 'fruit variety'
(d) DU' (Recursive DU):	[[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$T0]$_{DU'}$ → T2.T3-T2.T3.T0 *shui.guo-pin.zhong.de,* 'of fruit variety'
(e)IU:	{[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$}$_{IU}$ → T2.T3-T2.T3-T2.T3-T2.T3 *shui.guo-pin.zhong-zhi.you-ji.zhong,* 'There are only a few fruit varieties.'

Table 1. Formation of Different Levels of Prosodic Units.

In Table 1, the BU in (a) is formed on a Mandarin T3 sandhi domain. The BU in (b) forms a BU' (recursive BU) with T0 (the light tone), a clitic-like tone often carried by a handful of functional words. Two BUs in (c) form a typical DU on the higher level. The DU in (d) forms a DU' (recursive DU) with T0. Note that BU and DU differ only in size. Last, two DUs in (e) form a typical IU on the upper level.

The T3 sandhi applies not only in BUs as shown above but also across adjacent units as shown below:

(a) Sandhi applied to a compound structure:	[(T3.T3)$_{BU}$(T3)$_{BU}$]$_{DU}$→[T2.T3-T3]$_{DU}$→T2.T2-T3 *yu.san-chang,* 'an umbrella factory'
(b)Sandhi applied to a subject-predicate structure:	{[(T3.T3)$_{BU}$]$_{DU}$[(T3)$_{BU}$]$_{DU}$}$_{IU}$→{T2.T3-T3}$_{IU}$→T2.T2-T3 *yu.san-xiao,* 'The umbrella is small.'

Table 2. The Mandarin T3 Sandhi Applying across Adjacent Units.

[2] According to Li (2023), each full tone also has a head position, and the PU in this position indicates the tonal identity. For example, T3 as a dipping tone (L**L**H or 2**1**4) has the head position in the middle and occupied by **L** (bolded), indicating that it is essentially a low tone.

In Table 2, the phrase in (a) has a compounding structure, so the first two T3s and the last T3 can form two BUs respectively. Next, the two BUs form a DU. The T3 sandhi applies in the first BU and then across the two BUs. The clause in (b) has a subject-predicate structure, so two BUs can further form two DUs respectively. Then the two DUs form an IU. Likewise, the T3 sandhi applies in the first BU and then across the two DUs. The two examples indicate that the T3 sandhi is in essence prosody-based, but its applying order must make reference to morpho-syntactic structures.

Next, Shanghai Wu examples are presented to show how the proposed units are formed based on different types of tone sandhi. According to Li (2017), Shanghai Wu has prosodic words ranging from one to five syllables, and they usually form a tone sandhi domain to which a so-called "wide-application" type of sandhi applies. There is also a contrasting type called "narrow-application", occurring only at the phrase level. Each prosodic word as a wide-application sandhi domain forms a BU and each phrase as a narrow-application domain forms a DU. Therefore, Shanghai Wu tone sandhi is morpho-syntactically rather than prosodically based. The two types of sandhi domains are illustrated below:

(a)BU as the wide-application sandhi domain:	$(\underline{34}.\underline{13})_{BU} \rightarrow \underline{33}.\underline{44}$ *tshau.ve,* 'stir-fried cooked-rice'
(b)DU as the narrow-application sandhi domain:	$[(\underline{34})_{BU}(13)_{BU}]_{DU} \rightarrow \underline{33}$-13 *tshau-ve,* 'to stir-fry cooked-rice'

Table 3. Two Types of Sandhi Domains in Shanghai Wu.

In Table 3, the two-*zi* sequence *tshau ve* [3] carries 34 and 13 originally. In (a), the BU is associated with a prosodic word meaning 'stir-fried cooked-rice', to which the wide-application sandhi applies, turning 34.13 to 33.44. [4] In (b), each tone is associated with a prosodic word and two tones form two BUs respectively. Then the two BUs form a DU, associated with a verbal phrase meaning 'to stir-fry cooked-rice', to which the narrow-application sandhi applies, turning 34-13 to 33-13. The two resulting tone sequences are considered to have the left- and right-heavy prominence respectively (Guo, 2020). The linking relationship between different levels of units in the two Shanghai Wu examples can be further illustrated below:

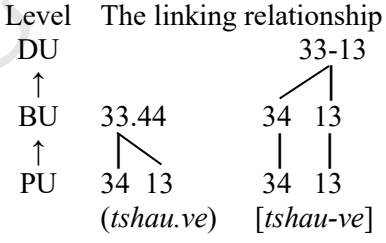


Figure 2. The Linking Relationship between Different Levels of Units in Shanghai Wu Examples.

In Figure 2, the vertical line is linked to a head position and the oblique line to a non-head position. The two-tone BU is linked to the head position on the left and branches to the right, as the wide-application sandhi results in the left-heavy prominence; The two-tone DU is linked to the head position on the right and

[3] Here the Wu spelling (https://www.wugniu.com) is used instead of the Mandarin Pinyin.
[4] 33 and 44 are not citation tones in Shanghai Wu. They only occur as a result of tone sandhi.

branches to the left, as the narrow-application sandhi results in the right-heavy prominence. Therefore, the linking relationship between different levels of units helps reflect prosodic prominence.[5]

4 Application of the proposed structure to an utterance

This section shows how the tone-based structure is further applied to a clause, "Once, the North wind and the Sun were disputing which was stronger", uttered by the same female speaker respectively in Mandarin and Shanghai Wu. First, the pitch contour of the Mandarin utterance is displayed below:

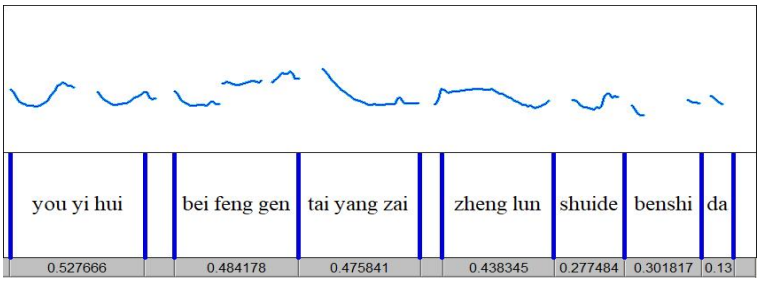


Figure 3. The Pitch Contour of the Mandarin Utterance.

In Figure 3, most units have a size of two or three tones. Also, the first few units are similarly timed, as evidenced by their time intervals around 0.5s. Note that the three-tone unit *you.yi.hui* ('once') has a little longer interval than the following two three-tone units, and the two-tone unit *zheng.lun* ('dispute') has a little shorter interval than the preceding three-tone units. This pattern suggests that both units are on the DU level by themselves, so the final lengthening effect applies to them, making the former longer than and the latter close in length to regular three-tone units.

In terms of melody, the whole contour lowers at the end, as evidenced by a large pitch height difference between the two high-falling T4s, one on *tai* in the word *tai.yang* ('sun') and the other on the final word *da* ('big'). Overall, the contour is relatively melodious. Now the underlying tone-based prosodic structure can be drawn for the Mandarin utterance as follows:

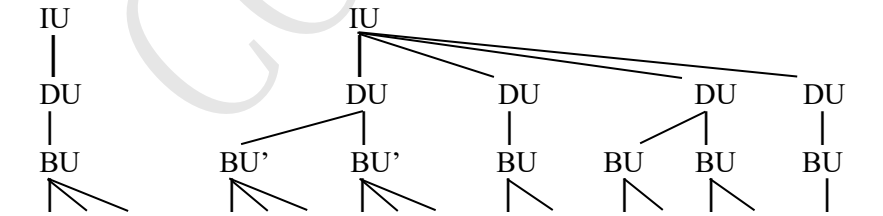


Figure 4. The Tone-based Prosodic Structure for the Mandarin Utterance.

In Figure 4, T1(55), T2(35), T3(214), and T4(51) are full tones and T0 is the light tone. Note that full tones can reduce to T0, a process called neutralization. For example, the high-rising T2 on *yang* in the word

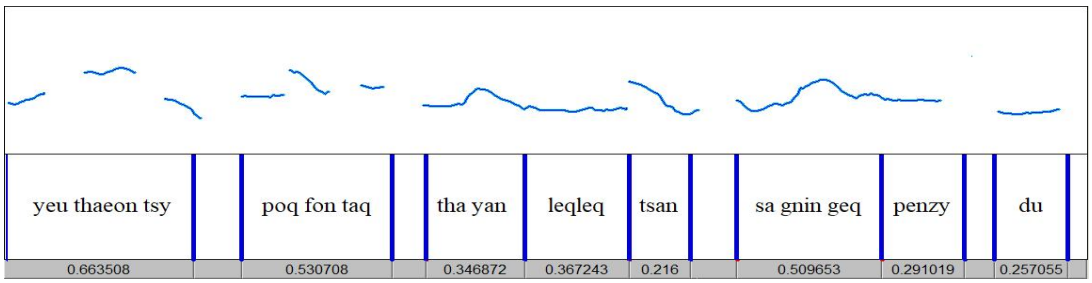
[5] Chen (2000) proposed two similar units, lexical and phrasal MRUs (Minimal Rhythmic Units), to account for different tone sandhi domains. However, he did not discuss their status in a prosodic hierarchy, let alone their linking relationship.

tai.yang (‘sun’) is usually neutralized to T0, as evidenced by a large fall instead of rise in the pitch contour shown in Figure 3. The first three tones can be considered as a single BU, because they are on the fixed expression *you.yi.hui* (‘once’). Furthermore, followed by the comma, this BU also forms a DU and an IU successively, as cued by final lengthening plus a relatively long pause (see the gap following *you.yi.hui* in Figure 3). Next, the two tones form a BU respectively in the words, *bei.feng* (‘north wind’), *tai.yang* (‘sun’), *zheng.lun* (‘dispute’), *shui.de* (‘whose’), and *ben.shi* (‘capability’). Then, T1 on *gen* (‘and’) cliticizes to the preceding BU on *bei.feng* (‘north wind’), together forming a BU’; likewise, T4 on the progressive aspect marker *zai* cliticizes to the preceding BU on *tai.yang* (‘sun’), together forming another BU’. Note that BU’s are placed on the same level as BUs for the sake of simplicity.

The two BU’s in the subject position further form a DU, as cued by a short pause (see the gap following *tai.yang.zai* in Figure 3). Next, the two-tone BU on the verb *zheng.lun* (‘dispute’) forms a DU by itself, as cued by final lengthening (see the long tail of its pitch contour in Figure 3). Then the following two BUs form another DU. Note that in the second BU *ben.shi* (T3.T4), the T4 syllable is much longer than the T3 one (in Figure 3, the large gap in the pitch contour belongs to the consonant *sh* in *shi*), indicating that it is lengthened at the phrasal boundary. Finally, the last tone on *da* (‘big’) forms a BU and then a DU successively due to its clause-final position. The last four DUs further form another IU in the utterance.

Also, a prosodic unit, if branched, is right-branching on the BU level, left-branching on the DU level, and right-branching on the IU level. What do these branching directions mean for Mandarin? First, the right-branching BU means that the head position must be on the left, and it cannot accommodate T0. T0 indeed occurs only on the right in a di-syllabic word, a non-head position. Also, if a monosyllabic functional word carries a full tone, the tone cannot be in the head position unless stressed. For example, T4 on the progressive aspect marker *zai* is actually neutralized to T0 just like the preceding T2 on *yang*, as evidenced by a large fall in the pitch contour shown in Figure 3. Therefore, both T0 and an unstressed tone can be treated as a clitic to the preceding tone or BU, rendering the BU or BU’ left-heavy. Next, the left-branching DU reflects the experimental finding that Mandarin phrases often use durational prominence, that is, final lengthening, to cue the right boundary, so they are considered to be right-heavy.[6] Last, the right-branching IU reflects the finding that Mandarin declarative sentences tend to lower the overall pitch level towards the end, hence rendering the IU left-heavy.

Next, the case with the Shanghai Wu utterance is presented, and its pitch contour is displayed below:



[6] Ma (2021) claimed that Mandarin has many compound words carrying a structural stress, whose placement largely depends on the context. The proposed structure places the left- and right-heavy prominence on two levels, suggesting that Mandarin word stress, if present, is more prosodic than lexical in nature, a view along the line with but more explicit than Ma’s (2021).

Figure 5. The Pitch Contour of the Shanghai Wu Utterance.

In Figure 5, most units are sized between two and three tones, but they are not as evenly timed as those in the Mandarin utterance, as evidenced by their varying intervals between 0.2s and 0.7s. Also, the whole contour does not lower much at the end, as evidenced by a small pitch height difference between the two low-rising T6s, one on *yeu* in the first phrase *yeu.thaeon.tsy* (‘once’) and the other on the final word *du* (‘big’). Overall, the Shanghai Wu utterance is less evenly timed and also less melodious than the Mandarin one. Now the underlying tone-based structure can be drawn for the Shanghai Wu utterance as follows:

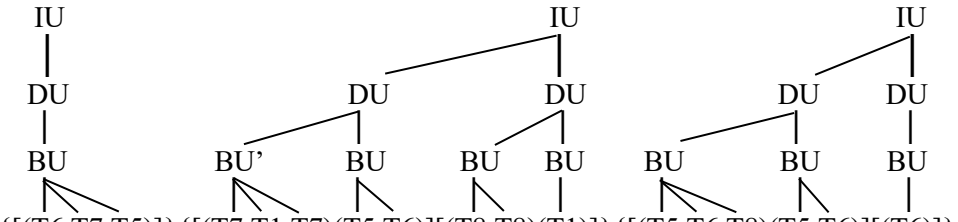


Figure 6. The Tone-based Prosodic Structure for the Shanghai Wu Utterance.

In Figure 6, T1(53), T5(34), T6(13), T7(5, glottalized), and T8(12, glottalized) are all citation tones. The prosodic structure of the Shanghai Wu utterance is clear-cut, since the wide-application sandhi domain is right-branching and left-heavy at the BU level and the narrow-application sandhi domain is left-branching and right-heavy at the DU level. Main structural features of the Shanghai Wu utterance are as follows: 1) following the monosyllabic verb *tsan* (‘dispute’), there is an extra IU boundary occurring with a relatively long pause (see the gap following *tsan* in Figure 5); 2) the progressive aspect marker *leqleq* forms a BU by itself and then a DU with the verb *tsan*; 3) the IU follows the DU in the branching direction and prominence position. They both are right-heavy, which conforms to the previous finding that Wu melody tends to have an overall small pitch fluctuation (Li, 2015).

The table below summarizes the structural and prosodic differences between the two types of Chinese:

Differences	Mandarin	Shanghai Wu
Tone sandhi type	T3 Sandhi/T→T0	Wide-/Narrow-Application
Branching direction	BU/DU/IU: Right/Left/Right	BU/DU/IU: Right/Left/Left
Pitch fluctuation(melody)	Large(more melodious)	Small(less melodious)
TPU length(timing)	More even(more TPU-timed)	Less even(less TPU-timed)

Table 4. Structural and Prosodic Differences between Mandarin and Shanghai Wu.

In Table 4, the first two rows indicate structural features and the last two prosodic ones. In the last row, the “TPU-timing” means that both BU and DU can be similarly timed. In the Mandarin utterance, BUs and DUs tend to have similar lengths when they contain the same number of tones, whereas in the Shanghai Wu one, they are less even in length. Therefore, the Shanghai Wu utterance is less TPU-timed than the Mandarin one. Note that in previous studies, Chinese rhythm is described in terms of syllable-timing. However, since the

proposed structure is tone-based, the description of Chinese rhythm needs to change from syllable- to tone-based accordingly. This change happens to support the notion of "tone-based rhythm" proposed by Xu (2018), and using TPU- instead of syllable-timing is a step forward to realize this notion.

5 A comparison with other structures

Hirst (2005) adopted a simple syllable-based four-level structure to capture the melodic and rhythmic differences between English and French. For English, the four levels of units from the bottom up are syllable (Syl.), rhythm unit (RU), tonal unit (TU), and intonation unit (IU). An RU can cover a trochaic stress foot and monosyllables outside the foot, delimited by word boundaries. A TU can cover RUs, delimited either by a foot or by an IU boundary. This structure allows a nuclear accent to occur non-phrase-finally, reflecting the left-heavy prominence of English ([H L]). For French, the four levels of units are Syl., TU, RU, and IU, with RU and TU reversed. TU being placed under RU indicates that no nuclear accent occurs phrase-initially, reflecting the right-heavy prominence of French ([L H]). Accordingly, English has a falling melody while French a rising one, and English has a stress-timed rhythm while French a syllable-timed one.

The advantage of this structure is its ability to integrate prosodic functions into a hierarchy without immediate morpho-syntactic concerns. However, this simple structure is not applicable to Chinese, because Chinese syllables are mostly toned, and the prosodic and morpho-syntactic interface is mainly made visible through tone sandhi rather than stress foot. Also, TU and RU, even if applicable to Chinese, can be consolidated into one given the pitch-timing duality of tone, and the TPU is just such one.

In addition, the tone-based structure is applicable to various Chinese dialects and able to provide a unified interface between prosody and morpho-syntax. This is because the prosodic features of different Chinese dialects can be reflected by the nature, head position, and branching direction of their TPUs. As long as tone sandhi patterns are identified, two levels of TPU (BU/DU) can be formed regardless of whether they are associated with words, phrases, or clauses. Now English, French, and Chinese are compared below in terms of their four-level structures and prosodic patterns. A simplified conventional structure is also included.

Prosodic structure	English	French	Chinese	Conventional
Level 4	IU	IU	IU	IP
	↑	↑	↑	↑
Level 3	TU	RU	DU	PP
	↑	↑	↑	↑
Level 2	RU	TU	BU	PW
	↑	↑	↑	↑
Level 1	Syl.	Syl.	PU	Syl.
Melodic pattern	[H L]	[L H]	[H L]/[L H]	Not Specified
Rhythmic pattern	Stress-timing	Syllable-timing	TPU-timing	Syllable-timing

Figure 7. A Comparison of English, French, Chinese, and Conventional Four-level Structures.

The last two rows in Figure 7 suggests that a syllable-based structure, however simple, is neither fully nor directly able to reveal Chinese melodic and rhythmic patterns as the tone-based one does. This is because Chinese can have the melodic pattern either falling or rising (corresponding to left- or right-heavy) at

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

different levels (hence [H L]/[L H]). Also, TPU-timing is a more appropriate term used to describe Chinese rhythm than stress- or syllable-timing, as it is entirely PU-based, and the PU number helps determine the tone and subsequent TPU length.

6 Implications of the proposed structure for prosody modeling

The above section shows that the proposed TP layer is related to different sizes or types of tone sandhi domains but does not directly refer to their morpho-syntactic status. In other words, TPU-based prosody processing does not need to rely heavily on grammatical markings, hence compatible with various theory-dependent models. However, any new model is subject to empirical validation, and the proposed one is no exception. In order to gain some insight for the future validation work, two prosody modeling studies are discussed below, because they are related to implications of the proposed structure.

First, Hirst (2015) developed a modeling tool, ProZed, based on the four-level English structure shown in Figure 7. This model is of theory-friendly nature and has an analysis-by-synthesis paradigm, hence allowing easy manipulation of rhythmic and tonal properties as well as immediate interactive experimentation with prosodic annotation. The proposed structure has a similar theory-friendly design, so it can be used to build a modeling tool to test different phonological theories on tonal behaviors. Better yet, it can accommodate a wide range of Chinese dialects with very different tone sandhi patterns and grammatical structures.

Second, Zu et al. (2022) discussed a so-called “sense element” (SE) in their experimental study of Chinese speech synthesis. The SE is a basic input unit encompassing both individual tones and multi-tone sandhi domains. Their results show that the SE-based input system leads to a higher MOS (Mean Opinion Score) than the traditional dictionary-based one in the assessment of the synthesized speech of Lhasa Tibetan. According to Zu et al. (2022), the main reason for SE to work well with Lhasa Tibetan data is its ability to maintain a good text and speech alignment, especially when functional words and verbal phrases are involved, both prone to phonology-syntax mismatch. When a mismatch occurs, the SE can be adjusted accordingly in size, since they are not confined to grammatical structures. The present study considers an SE in essence a BU or BU’, so it is expected that a BU-based system works as well. Nonetheless, the proposed structure is yet to be applied in future modeling projects and validated with favorable results.

7 Conclusion

The present study illustrates the tone-based prosodic structure and its application in two types of Chinese. Since this structure can accommodate both prosodically and morpho-syntatically based tone sandhi domains, there is no need to consider whether or not syllables are stressed: prosodic units are headed on each level and timed according to size. Unlike the above mentioned syllable-based structures, the proposed one not only helps capture melodic and rhythmic patterns of different Chinese dialects but also provides a theory-friendly framework for Chinese speech engineering.

References

Chengyu Guo. 2020. *Prosodic Structures and Tone Sandhi of Bi-syllabic Words in Chinese Dialects*. PhD Diss., Shanghai Normal University, Shanghai.

Daniel J. Hirst. 2005. Form and function in the representation of speech prosody. *Speech Communication,* 46(4): 334-347.

Daniel J. Hirst. 2013. Melody metrics for prosodic typology: comparing English, French, and Chinese. *Proceedings of Interspeech 2013*. 572–576.

Daniel J. Hirst. 2015. ProZed: A speech prosody editor for linguistic, using analysis-by-synthesis. In K Hirose, J Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*. Springer Verlag, Berlin Heidelberg. Chapter 1: 3–17.

Guhui Li. 2017. *On the Syntax-prosody Interface: Evidence from Tone Sandhi Processes in Chinese*. PhD Diss., East China Normal University, Shanghai.

Hongming Zhang. 2017. *Syntax-Phonology Interface: Argumentation from Tone Sandhi in Chinese Dialects*. Routledge, New York.

Huibin Zhuang. 2015. On defining *ci* in Chinese: A perspective from prosodic grammar. *TCSOL Studies*, 58(2):61–69.

Jianfen Cao. 2011. Prosodic structure and speech sound variation. *Journal of School of Chinese Language and Culture, Nanjing Normal University,* 3: 12–22.

Marina Nespor and Irene Vogel. 1986. *Prosodic Phonology*. Foris, Dordrecht, Netherlands.

Matthew Y. Chen. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge University Press, Cambridge, MA.

Peter Ladefoged and Keith Johnson. 2011. *A Course in Phonetics (Sixth Edition)*. Wadsworth, Boston, MA.

Qiuwu, Ma. 2021. Chinese word stress and its phonological properties. *Studies in Prosodic Grammar*, 8: 112–129.

Xi Ming Xu. 2018. *Contrastive Studies of Rhythmic Types in English and Mandarin.* Foreign Language Teaching and Research Press, Beijing.

Xiaobin Yan. 2010. A critical review of research on the phrasing of Chinese tone sandhi. *Journal of Nanjing University of Science and Technology (Social Sciences)*, 23(4): 76–83.

Ya Li. 2015. *Timing and Melody: An Acoustic Study of Rhythmic Patterns of Chinese Dialects.* PhD Diss., University of Victoria, Victoria, Canada.

Ya Li. 2023. Timing specification of Chinese tones. *International Journal of Chinese Linguistics*, 11(1): 76–97.

Yiqing Zu, Chen Lu, Ngodrup, Ronghua Zhu, Chenning Liu, Pengfei Shao, Klu' Bum Thr, Xiao Zhang, and Guoping Hu. 2022. Sense element in continuous speech: Evidence from Lhasa Tibetan speech synthesis. *Contemporary Linguistics*, 24(4): 515–532.

Youwei Shi. 2017. The dual-unit system in Chinese: Thoughts on the units of Chinese language teaching. *Journal of International Chinese Teaching*, 14(2): 34–41.

Zhigang Yin. 2020. Revisiting the acoustic characteristics and the generation mechanism of prosodic boundary. *Chinese Journal of Phonetics*, 13: 38–50.