INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Integrating the application and realization of Mandarin 3rd tone sandhi in the resolution of sentence ambiguity

Wei Lai, Aini Li

Department of Lingusitics, University of Pennsylvania
{weilai, liaini}@sas.upenn.edu

Abstract

Chinese third tone sandhi (T3S) covaries with the prosodic hierarchy both in the probability of application and in the realization of pitch slope. This paper evaluates whether Mandarin-speaking listeners integrate the covariation between T3S and prosody to resolute sentence ambiguity. Twenty-seven structurally ambiguous sentences were designed, each containing two consecutive T3 syllables situated across a word boundary, and the strength of the T3-intervening boundary crucially differentiates different interpretations of the sentence. The first T3 was manipulated to bear either a low, a shallow-rising, or a sharp-rising pitch. Sixty native Mandarin-speaking listeners heard each of these sentences and chose from two written interpretations the one that was consistent with what they heard. The results show that listeners are more likely to report a major-juncture interpretation when T3S does not apply (low) than when it applies (rising), and in the latter case, when the T3S variant has a sharper rather than shallower slope. Post-hoc analyses show that the T3S application is a more robust parsing cue for short sentences (4-5 syllables long), whereas the pitch shape of T3S is a more efficient parsing cue for longer sentences, indicating that listeners make sophisticated use of tonal variation to facilitate sentence processing.

Index Terms: Mandarin third tone sandhi, prosody, boundary, parsing, disambiguation

1. Introduction

Syntax and prosody are closely related and play a significant role in the computation of one another [1, 2]. A substantial body of empirical work has shown that listeners can exploit prosodic correlates of syntactic constituency to disambiguate structurally ambiguous sentences in auditory sentence processing [3, 4, 5]: When multiple interpretations are available for a sentence, listeners tend to prefer the analysis that is consistent with the provided prosodic information.

One way in which prosody informs sentence comprehension is through cueing the presence of a prosodic boundary that is crucial for sentence parsing [6, 7]. The most well studied phonetic manifests of prosodic boundary for guiding sentence parsing are pitch [8, 9] and duration [10, 11, 9]. In specific, listeners are sensitive to preboundary lengthening and pitch rising as cues for prosodic breaks. Moreover, they know which cues to attend to in the perception of prosodic boundaries at different levels [3, 12]. Other prosody-covarying cues than pitch and duration have been less attended to in psycholinguistic studies, including voice quality [13, 14, 15], consonant reduction [16, 17] and phonological variables with derivation domains licensed by prosodic structures [18, 19].

The present paper investigates a prosodically constrained phonological variable, namely, the third tone sandhi (T3S) in Mandarin Chinese, regarding its role in sentence disambiguation. Mandarin is a lexical tone language where fundamental frequency (F_0) is used to differentiate lexical meaning. According to the T3S rule, a low tone (T3) changes into a rising tone (similar to T2) when it precedes another T3. A solid body of work has investigated the domain where T3S is derived. The leading view regards T3S application as a diagnostic procedure of the prosodic domain of "foot", a prosodic construct defined by the binary-foot rule and the syntactic branching [20, 21]. For example, [20] proposed that the T3S application is obligatory within a foot, optional across foot boundaries, and prohibited cross intonation phrases delimited by pause and lengthening.

In spite of the rule-based aspects of T3S, quantitative research reveals a fine level of granularity in the probability and realizations of T3S. Although T3S is optional across feet for junctures smaller than intonation phrase boundaries, the probability of its application covaries with the prosodic structure in such a way that T3S is more likely to apply across smaller prosodic boundaries than larger ones [22]. Moreover, findings of acoustic measurements show that the pitch shape of T3S is distinct from that of T2, and its pitch realization varies with lexical and contextual factors such as word frequency [23] and positions in the prosodic hierarchy [22]. [22] demonstrated that within smaller prosodic domains where the two syllables are closely joined together and T3S obligatorily takes place, F_0 of the sandhi syllable rises sharply, but the duration is short, and its intensity is larger than its following T3 syllable; however, in larger domains where T3S optionally takes place, F_0 of the sandhi syllable rises smoothly with longer duration, and its intensity is weaker than the following T3 syllable.

The current paper evaluates whether Chinese listeners integrate variants of T3S to guide their parsing and arrive at different interpretations of the same ambiguous sentence in auditory sentence processing. This question contains a threefold inquiry:

1. Can listeners exploit the *probabilistic covariation* between T3S application and prosodic boundary, and arrive at boundary-consistent comprehensions with non-sandhied rather than with sandhied T3 variants?
2. Can listeners exploit the *gradient* covariation between the pitch slope of T3S variants and prosodic boundary, and arrive at boundary-consistent comprehensions with shallower rather than sharper T3S variants?
3. How does the effect of T3S, if any, interact with the integration of other well-known prosodic manifests such as phrasal lengthening in sentence disambiguation?

2. Method

2.1. Participants

Sixty participants were recruited to participate in an online experiment of Chinese auditory utterance comprehension through

http://dx.doi.org/10.21437/Interspeech.2020-2073

Qualtrics. They are 33 female 27 male, aged from 18 to 33 (mean = 23, sd = 3.6). All of them were reported to grow up in mainland China, speak Mandarin as their native language, and have no major hearing defects.

2.2. Stimulus design

The critical stimuli were 27 structurally ambiguous sentences, which are 4 to 8 syllables long. Each of these sentences contained two consecutive T3 syllables intervened by a word boundary. Different interpretations can be derived depending on whether the intervening boundary corresponds to a major or a minor syntactic juncture. Twenty-seven filler sentences (8-13 syllables long) without consecutive T3 syllables were included to increase variability. The filler sentences were also ambiguous but of different types.

2.2.1. T3S-dependent ambiguity resolution

The most frequent structure adopted in the critical stimulus sentences (22 out of 27 sentences) is [Verb NP_1 DE NP_2]. This structure is temporarily ambiguous between a complement-clause interpretation and a relative-clause interpretation [24], depending on whether there is a major syntactic juncture between VP and NP_1. Two T3 syllables were respectively placed in the VP-final and the NP_1-initial positions, so that listeners can infer the presence or absence of a boundary based on the sandhi information across the two T3 syllables. An example is given in (1).

(1) 逮捕 我们 的 间谍
dài bǔ wǒ men de jiàn dié
capture 1PL POSS/RC *spy*

(1a) [$_{\text{VP}}$ [$_{\text{VP}}$ dài bǔ] [$_{\text{NP}}$ wǒ men de jiàn dié]]
Arrest our spy.

(1b) [$_{\text{NP}}$ [$_{\text{CP}}$ dài bǔ wǒ men de] [$_{\text{NP}}$ jiàn dié]]
The spy that arrest us.

(1) can either mean “arrest our spy” with a major juncture between dài bǔ and wǒ men, as analyzed in (1a), or “the spy that arrests us” as analyzed in (1b), with no major-juncture between the T3 syllables. One way in which the two structures differ prosodically is that dài bǔ and wǒ men can be intervened with a large break (e.g., a long pause) in (1a), but not in (1b).

We manipulate the tones on bǔ to represent the conditions of no-sandhi, sharp-sandhi, and shallow-sandhi, and compare the boundary strength under each condition. If the application and realization of T3S matters, then it predicts that (1a) would be more frequently reported in the non-sandhi condition than in the sandhi condition, because T3S is more likely to apply within smaller prosodic domains. Similarly, it predicts that (1a) would be more frequently chosen in the shallow-sandhi condition relative to the sharp-sandhi condition, because the pitch shape of T3S becomes shallower across larger prosodic junctures.

2.2.2. T3S application and sentence length

Six out of the 27 sentences are not in the structure of [Verb NP_1 DE NP_2]. They are 4 to 5 syllables long, with relatively flatter syntactic and prosodic structures.

Presumably, these sentences allow us to examine the integration of the T3S application from a rule-based rather than probabilistic angle. In specific, sentences 4-5 syllables long tend to break down into combinations between two feet (2+2) or between a foot and superfoot (2+3, 3+2). In the meantime, T3S is *mandatory* within the domains of foot and superfoot, meaning that in short sentences, grouping between T3 syllables obligatorily causes T3S to apply. In this condition, the non-application of T3S is a strong cue of separation between the two T3 syllables. Consider the example in (2).

(2) 我 写 不 好
wǒ xiě bù hǎo
1SG *write* NEG *good/well*

(2a) [$_{\text{NP}}$ wǒ] [$_{\text{VP}}$ xiě bù hǎo]
I cannot write (it) well.

(2b) [$_{\text{CP}}$ wǒ xiě] [$_{\text{VP}}$ bù hǎo]
It is not good that I write (it).

In (2), T3S is mandatory under (2b) but optional under (2a). In other words, if T3S applies, the sentence is still ambiguous between the two interpretations; but if it does not apply, then the only interpretation available is (2a). This is different from longer sentences like (1) where the application of T3S is optional under both interpretations, but is more likely to take place under (1a). This difference makes the application of T3S a more robust cue of prosodic breaks in short sentences (4-5 syllables) than in long sentences (more than 5 syllables).

2.3. Recording and manipulation

The stimulus and filler sentences were recorded in a professional recording booth by a female Mandarin speaker (the first author). The 27 stimulus sentences were repeated 6 times with sandhi and 6 times without sandhi; in each of the two conditions, three repetitions were read with a timing pattern in favor of a major-juncture interpretation, and the other three times were in favor of a minor-juncture interpretation. The filler sentences were read twice with two prosodic patterns in favor of different interpretations.

For each sentence, one well-articulated sandhied repetition was then chosen, and their T3 syllables were manipulated in pitch and timing. In pitch manipulation, the T3 syllables were manipulated into three tone conditions: sharp-rising, shallow-rising, and low. The resulting pitch contours of the three tone conditions are shown in Fig. 1.

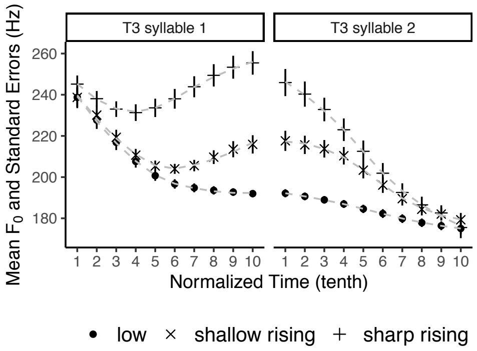

Figure 1: *Time-normalized F_0 contours of the two T3 syllables under the three tone conditions*

Then the 27 sentences with three tone conditions were each manipulated into two timing conditions: normal and shortened. We forced-aligned all the recordings using the Penn Phonetics Lab Forced Aligner [25], and adjusted any misaligned segment boundaries by hand. Then, the mean duration of each segment in each sentence was measured by averaging across its 12 repetitions. For each sentence, a normal-timing version was generated by scaling the duration of each segment in the sentence to the grand mean of itself, and a temporally shortened version was then generated by compressing the first T3 syllable of its normal-timing counterpart to 0.7 of its original length. In each timing condition, stimuli of the same sentence with different tone realizations still share the same duration setting.

2.4. Procedure

Three experimental lists were constructed, each made up of 54 experimental stimuli and 46 filler items. The 54 experimental stimuli were 9 low-tone items, 9 sharp-rising sandhi items, and 9 shallow-rising sandhi items, each occurring twice in different temporal conditions. The pairing of items with tone conditions was counterbalanced across lists, so that each participant heard each experimental stimulus in only one of the three tone conditions, and heard them twice in both shortened and normal temporal conditions. Participants were randomly assigned to one of three lists, with an equal number of participants (N = 20) hearing each list.

Listeners were told that they would hear sentences that were ambiguous in meaning. For each sentence, they would see two options of interpretations, each paraphrased in unambiguous ways. Listeners were asked to identify the interpretation that was consistent with the speech they heard. Each listener did three practice trials with filler sentences before they proceeded to the experiment. All of the one hundred trials (54 critical stimuli and 46 fillers) were presented in a single block, with the order of trials and the choices of each trial randomized for each participant.

3. Results

The aggregate major-juncture reading rates on each tone and timing conditions are illustrated in Fig. 2. It shows that the frequency of major-juncture reading varies with both timing and tone. The major-juncture reading rate decreases as the duration of the first T3 syllable becomes shorter. Within each timing condition, the likelihood of a major-juncture reading is the highest in the low condition, becomes lower in the shallow-rising condition, and reaches the lowest in the sharp-rising condition.

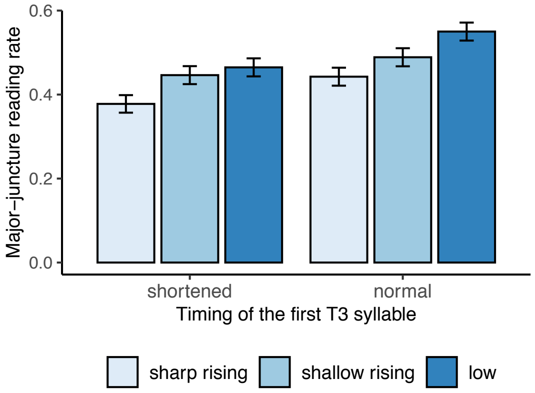

Figure 2: *The means and standard errors of the major-juncture reading rate under different tone and timing conditions*

A mixed-effects logistic model was conducted to predict the Response of a major-juncture reading, with Tone-nested-in-Timing as the fixed effect (Timing: sum-coded, shortened: 1, normal: -1; Tone: repeated coded, sharpR-shallowR-Low), and Sentence and Subject-nested-in-Group as random intercepts. The results, as shown in Table 1, reveal an overall significant Timing effect (β = -0.16, p<0.001), indicating that the log-odd of a major-juncture reading is 0.16 higher in the shortened condition than the normal condition. With normal timing, the log probability of a major-juncture reading was significantly higher in the Low condition than the shallow-rising condition (β=0.3, p=0.02) and marginally higher in the shallow-rising conditions than the sharp-rising condition (β=0.23, p=0.09). With shortened initial T3, the major-juncture rate shows no difference between the shallow-rising and the low condition (β= 0.09, n.s.), and the rate is significantly higher in the shallow-rising than the sharp-rising condition (β= 0.35, p=0.01). These results indicate that listeners are sensitive to the phonological application of T3S (low versus rising), as well as the pitch slope of the T3S variant (sharp rising versus shallow rising).

Following Section 2.2.2, we divided all the sentences into short and long with the threshold of 5 syllables, and examined the potential effect of sentence length. The major-juncture reading rates for short (4-5 syllables) and long sentences (more than 5 syllables) are presented in Fig. 3.

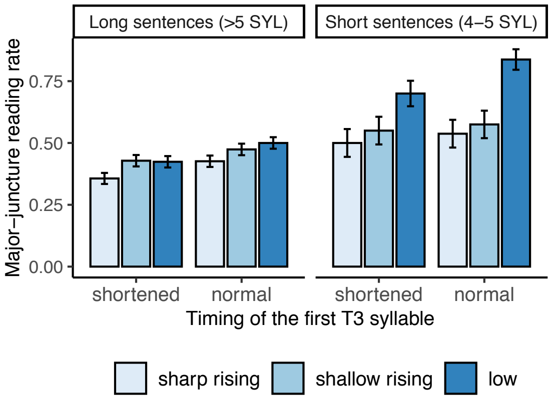

Figure 3: *The means and standard errors of the major-juncture reading rate for long and short sentences*

The mean major-juncture rate shows very different patterns depending on sentence length. In short sentences, the application of T3S plays a significant role, such that the major-juncture rate increased from 0.5-0.6 in the sandied conditions (sharp-rising, shallow-rising) to 0.7-0.8 in the non-sandhi condition (low). In long sentences, however, the pitch slope of the T3S variant seems to matter more than the T3S application, such that the difference between the two rising conditions becomes larger than the difference between the shallow-rising and the low conditions. The large standard errors in the right facet result from the relatively small number of short sentences (N=4), but the pattern still appears to be quite robust.

As with the aggregate result, two mixed-effects logistic models were estimated, one for sentence length (short, long), with the results shown in Table 1, along with the aggregate result. It shows a significant difference in the major-juncture reading rate between the low and the shallow rising conditions for short sentences with normal timing (β=1.44, p =0.001), and between the two rising conditions for long sentences with shortened timing (β=0.36, p =0.015). In addition, a general timing effect is found for both short (β=-0.20, p =0.06) and long (β=-0.16, p <0.001) sentences. These results suggest that the cue of T3S application is more useful for the disambiguation of short sentences, whereas the pitch slope of T3S variants is more use-

Table 1: *Results of GLM* $Response \sim Timing/Tone + (1|Group/Subj) + (1|Sentence)$ *for different sentence lengths*

Sentence type	**All sentences**			**Short sentences**			**Long sentences**		
Fixed effects	β	SE	z	β	SE	z	β	SE	z
(Intercept)	-0.16	0.20	-0.82	0.59	0.38	1.54	-0.28	0.20	-1.37
timing(shortened)	**-0.16**	**0.04**	**-4.06*****	-0.20	0.11	-1.89.	**-0.16**	**0.04**	**-3.75*****
timing(normal):tone(Low-shallowR)	**0.30**	**0.13**	**2.25***	**1.44**	**0.44**	**3.29****	0.13	0.15	0.93
timing(normal):tone(shallowR-sharpR)	0.23	0.13	1.68.	0.38	0.47	0.81	0.23	0.15	1.57
timing(shortened):tone(Low-shallowR)	0.09	0.13	0.69	0.65	0.40	1.64	-0.01	0.15	-0.09
timing(shortened):tone(shallowR-sharpR)	**0.35**	**0.14**	**2.54***	0.44	0.47	0.93	**0.36**	**0.15**	**2.43***

ful for the disambiguation of long sentences.

4. Discussion and Conclusion

This paper uses sentence disambiguation as a testing ground to evaluate the hypothesis that listeners of Mandarin integrate the probabilistic and gradient properties of third tone sandhi into sentence processing. The result shows different major-juncture reading rates depending on whether the first of two consecutive T3 syllables bears a rising or low pitch, indicating that listeners' sentence parsing decisions are affected both by whether third tone sandhi applies or not. It also shows different major-juncture reading rates between sentences with a sharp-rising variant and those with a shallow-rising variant, indicating that when tone sandhi applies, listeners' boundary perception is sensitive to the pitch slope of the T3S variant. The integration of T3S further interacts with other prosodic influences, such that an additional timing effect was observed across tone conditions that compressed duration of the first T3 syllable would lead to a lower major-juncture reading rate overall. These results indicate that listeners are able to exploit both probabilistic and phonetic information of tone variants in sentence processing.

Note that not all the empirically observed differences turn out to be statistically significant. For example, no difference in the major-juncture comprehension rate was found between the low condition and the shallow-rising condition with a shortened first T3 syllable. We think that this lack of difference can be attributed to how the stimuli were constructed. In the manipulation of shortened-timing stimuli, time compression was implemented *after* pitch alteration, and no further pitch adjustment was conducted afterward. As a result, temporally shortened shallow-rising variants have achieved the same amount of pitch change within a shorter time period, ending up bearing a somewhat dipping pitch contour with steeper falling and rising components. We expect this to cause a perceptual asymmetry of the shallow-rising variant between different timing conditions.

The effect of sentence length was not anticipated but turned out to be linguistically justifiable. As elaborated in Section 2.2.2, sentences that are 4-5 syllable long either consist of two binary feet (2+2) or one binary foot and one superfoot (2+3/3+2), making T3S mandatory under the interpretations where the two T3 syllables are grouped together. This predicts that for short sentences, the lack of tone sandhi strongly entails a major-juncture reading, whereas such a bias does not exist for sandhied stimuli. In contrast, in longer sentences, T3S is optional under both interpretations and only differ in probability. An astonishing finding is that listeners are able to make use of the division of cue efficiency between short and long sentences for the same tone-prosody covariation. Their responses reflect a heavier influence of the phonological cue of T3S application for short sentences, as indicated by the large primary contrast between low and rising tone conditions in the distribution of responses, and a heavier influence of the phonetic cue of pitch shape for long sentences, as indicated by the major difference in response distributions across the two rising tone conditions with different pitch slopes.

The above results have broader implications on listeners' knowledge of speech variability originating in phonetic interpolation and phonological derivation. In particular, the finding that listeners shift their reliance between the phonological application and the phonetic realization of T3S depending on sentence length suggests that listeners keep track of the association between speech variability and contextual factors in mutually orthogonal dimensions. These results lend support to such an account: Listeners acquire nuanced knowledge of speech variation and their distribution in dispersed phonological and phonetic contexts through extensive native language exposure. They make efficient use of this knowledge in speech processing by integrating the most relevant aspect(s) of speech variability as appropriate to the impending context.

Inevitably, the stimulus sentences also have their intrinsic biases towards one interpretation or the other that are attributable to the structural and lexical properties of the sentence, rather than the acoustic realizations. While we do not consider this to undermine our finding of the difference between conditions, since the sentences are counterbalanced across conditions, it is possible that another set of stimuli might have widened the gap between conditions to a larger extent. It is also worth noting that the aggregate pattern masks considerable individual variability, suggesting that Not all the participants have shown a trend consistent with the aggregate pattern. It remains a question how individual listener's strategies are related to their language background and experience. Future work can pursue these questions.

5. References

[1] E. Selkirk, "Phonology and syntax: the relation between sound and structure," *Current Studies in Linguistics I*, 1984.

[2] M. Nespor and I. Vogel, "Prosodic structure above the word," in *Prosody: Models and measurements*. Springer, 1983, pp. 123–140.

[3] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *the Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[4] W. D. Marslen-Wilson, L. K. Tyler, P. Warren, P. Grenier, and C. S. Lee, "Prosodic effects in minimal attachment," *The Quarterly Journal of experimental psychology*, vol. 45, no. 1, pp. 73–87, 1992.

[5] A. J. Schafer, S. R. Speer, P. Warren, and S. D. White, "Intonational disambiguation in sentence production and comprehension," *Journal of psycholinguistic research*, vol. 29, no. 2, pp. 169–182, 2000.

[6] S. R. Speer, M. M. Kjelgaard, and K. M. Dobroth, “The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities,” *Journal of psycholinguistic research*, vol. 25, no. 2, pp. 249–271, 1996.

[7] J. Pynte, “Prosodic breaks and attachment decisions in sentence parsing,” *Language and cognitive processes*, vol. 11, no. 1-2, pp. 165–192, 1996.

[8] C. M. Beach, “The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations,” *Journal of memory and language*, vol. 30, no. 6, pp. 644–663, 1991.

[9] L. A. Streeter, “Acoustic determinants of phrase boundary perception,” *The Journal of the Acoustical Society of America*, vol. 64, no. 6, pp. 1582–1592, 1978.

[10] I. Lehiste, “Phonetic disambiguation of syntactic ambiguity,” *The Journal of the Acoustical Society of America*, vol. 53, no. 1, pp. 380–380, 1973.

[11] D. R. Scott, “Duration as a cue to the perception of a phrase boundary,” *The Journal of the Acoustical Society of America*, vol. 71, no. 4, pp. 996–1007, 1982.

[12] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, “Segmental durations in the vicinity of prosodic phrase boundaries,” *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[13] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, “Glottalization of word-initial vowels as a function of prosodic structure,” *Journal of phonetics*, vol. 24, no. 4, pp. 423–444, 1996.

[14] H. Zhang, “Boundary effects on allophonic creaky voice: A case study of mandarin lexical tones,” *Tonal Aspects of Languages 2016*, pp. 94–98, 2016.

[15] J. Kuang, “Creaky voice as a function of tonal categories and prosodic boundaries.” in *INTERSPEECH*, 2017, pp. 3216–3220.

[16] J. Yuan and M. Liberman, “Investigating consonant reduction in mandarin chinese with improved forced alignment,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] W. Lai and J. Kuang, “Prosodic grouping in chinese trisyllabic structures by multiple cues–tone coarticulation, tone sandhi and consonant lenition,” *Tonal Aspects of Languages 2016*, pp. 157–161, 2016.

[18] J. Itô, “A prosodic theory of epenthesis,” *Natural Language & Linguistic Theory*, vol. 7, no. 2, pp. 217–259, 1989.

[19] H. Van der Hulst, “Vowel harmony,” in *Oxford Research Encyclopedia of Linguistics*, 2016.

[20] C.-l. Shih, *The prosodic domain of tone sandhi in Chinese*. University of California, San Diego, 1986.

[21] M. Y. Chen, *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press, 2000, vol. 92.

[22] J. Kuang and H. Wang, “T3 sandhi at the boundaries of different prosodic hierarchies,” *J. Chinese Phonetics*, vol. 1, pp. 125–131, 2006.

[23] J. Yuan and Y. Chen, “3rd tone sandhi in standard chinese: A corpus approach,” *Journal of Chinese Linguistics*, vol. 42, no. 1, pp. 218–237, 2014.

[24] Y. Hsieh, J. E. Boland, Y. Zhang, and M. Yan, “Limited syntactic parallelism in chinese ambiguity resolution,” *Language and Cognitive Processes*, vol. 24, no. 7-8, pp. 1227–1264, 2009.

[25] J. Yuan and M. Liberman, “Speaker identification on the SCOTUS corpus,” *Proceedings of Acoustics 2008*, pp. 5687–5690, 2008.