EFFECT OF CONTEXTUAL TONAL VARIATION ON SPEECH RECOGNITION: EVIDENCE FROM EYE MOVEMENTS

Qian Li[a]; Yiya Chen[a,b]

[a] Leiden University Center for Linguistics; [b] Leiden Institute for Brain and Cognition
q.li@hum.leidenuniv.nl; yiya.chen@hum.leidenuniv.nl

ABSTRACT

While much is known about how lexical tones are perceived in isolation, little has been done on the perception of tones in connected speech. This study, via visual world paradigm, investigates effect of contextual tonal variation on speech recognition in Tianjin Mandarin, where three types of contextual tonal variation have been identified: Near-Merger Sandhi, No-Merger Sandhi, and No-Sandhi Coarticulation. Listeners were asked to identify the target amongst an array of four possibilities upon hearing a disyllabic collocation, while their eye movements were tracked. Results suggest that native listeners are sensitive to fine-grained phonetic details in contextual variation of lexical tones. No-Sandhi Coarticulation was the easiest to recognize as participants fixated on the targets the earliest among three conditions. Near-Merger Sandhi was more difficult to process than No-Merger Sandhi, reflected in the overall less proportion of looks to target.

Keywords: tonal variation, tone perception, Tianjin Mandarin, eye tracking, Visual World Paradigm

1. INTRODUCTION

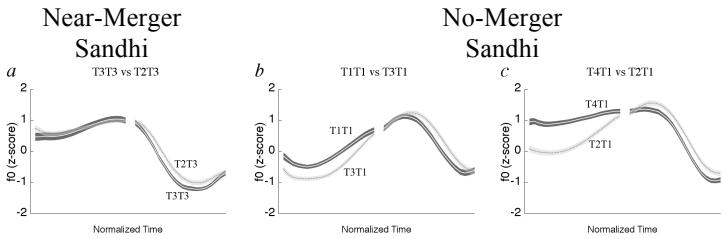
In many tonal languages, tones are primarily signalled via different *f0* patterns [7]. Studies have revealed that native listeners of tone languages utilize various types of *f0* cues to identify lexical tones [e.g., 5, 6, 13, 16, 18, 19, 20]. These findings have shown the importance of *f0* information in identifying the canonical realization of lexical tones produced in isolation. Little, however, is known on how tones are perceived in connected speech.

In connected speech, the *f0* contours of lexical tones may deviate greatly from that produced in isolation, due to contextual effects such as tone sandhi and tonal coarticulation ([4, 24] and references therein). Tone sandhi typically refers to phonological tonal alternations, which usually causes greater changes in the *f0* contour so that sandhi-derived tones are very often unpredictable or unexpected from the canonical *f0* realization. A well-known case is the Low tone sandhi in Beijing Mandarin. When two Low tones are combined in a disyllabic domain, the first Low (which otherwise would be realized with a low *f0* target) surfaces with a rising *f0* contour ([3] and references therein). Note that even for tones without tone sandhi allophonic variation, they are usually realized with an *f0* contour that is different from their canonical *f0* shape. The deviation, however, is predictable and much subtler with varying degrees of phonetic modification as a function of the tonal context. This is known as tonal coarticulation. Despite the extensive contextual variation in the *f0* realization of lexical tones in connected speech and the rich layers of different types of tonal variation, only a small number of perception studies have taken the contextual effect into consideration [e.g., 5, 6, 16]. This body of work shows consistently that correct identification of lexical tones relies on tonal context when the *f0* contours deviate greatly from the canonical *f0* realization. However, the time-course of processing contextual tonal variation during on-line speech recognition remains unknown. In the present study, we set out to address the question via investigating the processing of tonal variation in Tianjin Mandarin (TM) with a speech recognition task within the visual world paradigm [21].

TM presents a good testing case as it exhibits interesting patterns of tonal variability in connected speech. There are four lexical tones in TM: Tone 1 (T1) is a low-falling tone, Tone 2 (T2) a high-rising tone, Tone 3 (T3) a dipping tone, and Tone 4 (T4) a high-falling tone [11, 25]. In connected speech, TM shows a range of complex tone sandhi patterns over disyllabic constituents. These sandhi patterns have been claimed to involve the categorical change of one lexical tone to another. For example, when two T3s are combined, the first one is claimed to change into T2 [e.g., 3, 9, 12]. However, recent experimental data have shown that there is no complete neutralization in TM and sandhi does not involve the categorical change of one lexical tone to another [10, 25]. Furthermore, two different types of tone sandhi have been proposed in this language, one is Near-Merger Sandhi and the other No-Merger Sandhi, as illustrated in Figure 1 [11].

Figure 1: *f0* realization of three disyllabic sandhi sequences in TM (T3T3 in *a*, T1T1 in *b* and T4T1 in *c*) compared to their respective targeted patterns claimed in the literature. T3T3 is Near-Merger

Sandhi; T1T1 and T4T1 No-Merger Sandhi. Lines indicate the mean *f0;* shades for ±1 SE. Normalized time.



In the Near-Merger Sandhi (i.e., T3T3, Figure 1*a*), the first T3 is realized with a high-rising *f0*, which makes the *f0* realization of T3T3 hardly distinguishable from that of the claimed target sequence (T2T3) in the literature. In the No-Merger Sandhi, the sandhi tones surface with altered *f0* realization, while maintaining its distinctiveness from the other lexical tonal contours. There are two No-Merger Sandhi cases in TM: T1T1 (Figure 1*b*) and T4T1 (Figure 1*c*). T1T1 has been claimed to change into T3T1 in the literature [e.g., 3, 9, 12]. As in Figure 1*b*, the first T1 in T1T1 is realized with a slightly falling and rising *f0*, which is unexpected from its canonical low-falling *f0* when produced in isolation. Although the sandhi-derived T1 is realized similarly to T3 as in T3T1, its *f0* realization is significantly different from that of T3, suggesting no merge of the sandhi-derived T1 with T3. Similarly, T4T1 is claimed to change into T2T1 [e.g., 3, 9, 12]. While the first sandhi-derived T4 in T4T1 is similar to the T2 contour, they are clearly different from each other (Figure 1*c*). Given the range of variability, the specific research question of interest here is whether and how the different types of tone sandhi variation affect tonal processing in online speech recognition?

2. METHOD

2.1. Participants

31 native speakers of TM participated in this experiment. All were born in late 1980s to early 1990s and raised in the urban areas of Tianjin. They were undergraduate or postgraduate students studying in Beijing at the time of experiment. None of them had lived out of Tianjin before 18. They were paid for participation but unaware of the purpose of the experiment. All had normal or corrected-to-normal vision.

2.2. Stimuli

Target stimuli consisted of 36 highly lexicalized disyllabic collocations with two sandhi patterns in TM: Near-Merger Sandhi (i.e., T3T3) and No-Merger Sandhi (i.e., T1T1, T4T1). In addition, 18 stimuli without sandhi changes were included for further comparison (i.e., T4T4, T3T2, T3T4).

For each target stimuli, a corresponding baseline competitor and a critical competitor were chosen. The baseline competitor did not share segment or tone with the target (S-, T-). The critical competitor shared the segment of the first syllable with the target stimuli, but with an unrelated tone (S+, T-), which is neither the underlying lexical tone nor the lexical tone that is claimed in the literature as the targeted sandhi tone. The second syllables of the target and competitor were different in terms of both tone and segment. Each target-competitor pair also had two distractors within the visual-world paradigm. Distractors did not share any tone or segment with the target or the competitor.

Targets and competitors were closely matched in terms of frequency based on [2] and orthographic complexity as we presented Chinese characters instead of pictures (following [20]), so that there was no significant difference between the targets and competitors for frequency ($F=0.087$, $p=0.768$) or visual complexity ($F=0.156$, $p=0.694$).

Target stimuli were pre-recorded by a male speaker of TM who was born in the 1980s. All auditory stimuli were produced with the same loudness and speaking rate. The mean duration of the first syllable of the stimuli was 357ms (SD=71ms), and not significantly different across three conditions ($F=0.47$, $p=0.63$).

2.3. Procedure

Eye movements were recorded in Eyelink 1000 with a 35mm lens running at 500Hz. Visual stimuli were presented on a 21-inch monitor. Participants were seated comfortably with a chin rest and a forehand rest set at a distance of 69cm from the screen.

The experiment began with seven training trials to familiarize the participants with the experimental setup and task. Participants were tested individually. Each trial consisted of a fixation cross screen with only a fixation cross at the screen center (500ms) followed by a stimuli screen. The participants were asked to look at the cross until it disappeared. Then the auditory stimulus was played through a headphone simultaneously with the presentation of the stimuli screen. The task was to click on the word they had heard with a mouse.

2.4. Data Analysis

The proportion of looks at target at each time point (every 2ms) was calculated. Trials in which subjects clicked on items other than the target (<1%) were

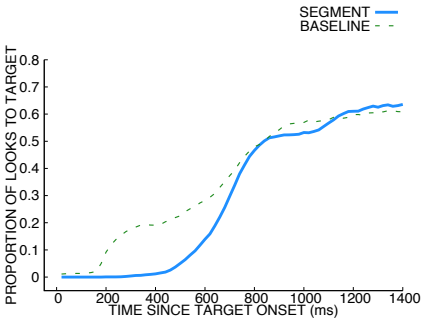
excluded from the analysis. In addition, for a better illustration of the eye movement patterns, data containing blinks (17%) were also excluded. We followed the growth curve analysis procedure introduced in [15] for statistical analyses with the package *lme4* [1] in R [17].

3. RESULTS

3.1 Baseline Comparison

Figure 2 illustrates the proportion of looks to target averaged across three tonal variation types when the competitor is a baseline competitor (i.e., without segmental overlap) vs. when it is a critical competitor (i.e., with segmental overlap). X-axis stands for the time since target onset and the y-axis for the proportion of looks to target.

Figure 2: Proportion of looks to target over time when the target is presented with a critical competitor (SEGMENT) vs. a baseline competitor (BASELINE).

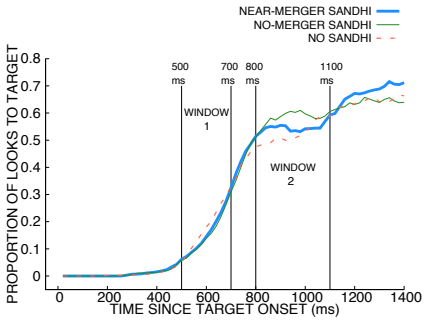


As suggested by Figure 2, when there is segment overlap between targets and competitors (solid lines as SEGMENT), the proportion of looks to target is clearly smaller in the time window of 200-800ms than where there is a baseline competitor (dash lines as BASELINE). The two conditions showed a similar gaze pattern after 800 ms from the stimuli onset. A growth curve analysis was run to compare the SEGMENT and BASELINE conditions within the 200-800ms time window. Results showed a significantly different overall mean, slope and contour shape between the two conditions (*intercept*: Est.=0.15, t=33.44, p<0.001; *slope*: Est.=-0.18, t=-7.38, p<0.001; *quadratic*: Est.=-0.13, t=-5.34, p<0.001; cubic: Est.=0.09, t=3.86, p<0.001).

3.2 Different Types of Contextual Tonal Variation

Figure 3 shows the proportion of looks to target for three different tonal variation types when the targets and competitors have segment overlap. X-axis stands for the time since the auditory target onset, y-axis for the proportion of looks to target.

Figure 3: The proportion of looks to target when target and competitor have segment overlap in three tone sandhi types, aggregated across participants and items.



In Figure 3, there are subtle differences in the eye-movement patterns due to different types of tonal variation in the target stimuli. The proportion of looks to target in all tonal variation types remains at the bottom (less than 0.1) until around 500ms when the proportion starts to rise. After 500ms, the proportion of looks to target in different tone sandhi types begins to diverge, especially in two time windows: 500-700ms and 800-1100ms.

In the 500-700ms window (WINDOW 1, Figure 3), the No-Sandhi condition had a significantly higher overall mean of the proportion of looks to the target than both Near-Merger Sandhi (*intercept*: Est.=-0.02, t=-2.48, p=0.01) and No-Merger Sandhi (*intercept*: Est.=-0.02, t=-3.53, p<0.001), while the two sandhi conditions were not significantly different. In addition, the significant difference in the second order time term suggested that the No-Sandhi condition had a steeper rising of the proportion of looks to target than the two sandhi conditions (*slope* Near-Merger: Est.=0.05, t=1.99, p<0.05; No-Merger: Est.=0.05, t=2.03, p=0.04).

The second time window of interest is between 800-1100ms (WINDOW 2, Figure 3), where there are noticeable differences among the three conditions. Two main observations can be made. First, the No-Sandhi condition differred from the two sandhi conditions in terms of the overall mean proportion of looks to target. There was significantly less proportion of looks at the target in the No-Sandhi condition than both the Near-Merger Sandhi (*intercept*: Est.=0.02, t=2.80, p=0.005) and the No-Merger Sandhi (*intercept*: Est.=0.07, t=8.90, p<0.001). Second, there were overall less looks to the target in Near-Merger Sandhi than that in the No-Merger Sandhi condition (*intercept*: Est.=0.05, t=6.01, p<0.001).

4. DISCUSSION & CONCLUSION

We employed the Visual World Paradigm to investigate the time course of disyllabic tone perception in TM. Our results suggest significant perceptual differences among different types of tone variation, which affects on-line speech processing differently, as reflected in the different eye movement patterns.

First, we made the baseline comparison where the target and competitor had neither segmental nor tonal overlap. We observed a generally smaller proportion of looks to the target when there was segment overlap than that in the baseline condition. This indicates a stronger competition effect between target and competitor when there is segmental overlap. Our results showed a comparable pattern to a recent study on monosyllabic Mandarin tone perception [14]; when target and competitor have segment overlap and tonal mismatch, the proportion of looks to target in the two conditions started to diverge around 600ms post target onset. In addition, our data showed much earlier looks to the target in the baseline condition (around 200ms). This is probably due to the fact that we used printed words as visual stimuli, which is proposed to be more sensitive to phonological manipulations than images for alphabetic languages [8, 22]. Further studies are needed to verify the potentially different effects of image vs. printed words on gaze patterns in logographic languages during speech processing.

We also compared the proportion of looks to targets with three different tonal variation types when target and competitor have segmental overlap. The No-Sandhi condition showed quicker increase of looks to the target compared to two sandhi conditions in the time window of 500ms to 700ms. In the time window of 800-1100ms, participants did not seem to look at the No-Sandhi targets any more than what they did in two sandhi conditions. This might be due to that, in the No-Sandhi condition, tonal processing is relatively easier given that the tones are realized with only subtle phonetic deviations. Our analyses of looks to competitors also suggested that participants by then have already shifted their attention from both targets and competitors to somewhere else, probably due to the ease of processing. Comparatively, the two sandhi conditions present more challenges for tonal recognition due to the more dramatic *f0* contour distortion. This is reflected in their delayed fixations on target. Between the two sandhi conditions, we observed different eye movement patterns in the time window from 800ms to 1100ms. Looks to target showed significantly higher proportion for No-Merger Sandhi than Near-Merger Sandhi, which suggests that tonal identification is relatively easier for stimuli with No-Merger sandhi than with Near-Merger Sandhi.

Previous studies on tone perception in connected speech showed a clear effect of tonal contexts on lexical tone identification [e.g., 5, 6, 16]. Tonal context is especially important for tone perception when tones deviate greatly from their canonical realizations [23]. Our results show, for the first time, that tonal variation within the first syllable has already exerted an effect on tonal recognition, given comparable contextual facilitation from the second syllable for tonal identification. First, for the No-Sandhi condition, where there is only minor f0 deviation with distinctive lexical tonal contours maintained, listeners clearly relied relatively less on the contexts and identified the target tones earlier. Second, for the two sandhi conditions, where the *f0* realization of the lexical tones is altered to such a great extent that tonal distinctiveness is no longer kept, it is difficult to identify the target tone only based on the sandhi-derived f0 contours. Information of the second syllable was thus needed to a greater extent in the identification of the lexical tone of the first syllable. Between the two sandhi conditions, the lower proportion of looks to target in Near-Merger Sandhi suggested more difficulty because the sandhi-derived tone is almost identical to another tone within the tonal system, resulting more competition in recognition, compared to No-Merger Sandhi where the sandhi-derived tones remained quite distinctive from even the most similar lexical tone within the tonal inventory.

To summarize, our data suggest that tone sandhi is processed differently from no sandhi tonal coarticulation. In addition, there is also the need to differentiate two different types of tone sandhi in TM, which clearly have different consequences on tonal processing. The implication of our results on possible linguistic theory of tonal variability and sandhi alternations will be explored in the future.

5. ACKNOWLEDGEMENT

This research was supported by a CSC scholarship to QLi, and an ERC grant to YChen (ERC-206198). We thank Ting Zou for stimuli recording and piloting. Thanks to Wei Zhou and Ming Yan for making the eye-tracker available at Beijing Normal University. Special thanks to Kurt Debono from SR Research for the timely technical support.

6. REFERENCES

[1] Bates, D., et al. 2014. lme4: Linear Mixed-effects Models using Eigen and S4. http://cran.r-project.org/web/packages/lme4/index.html

[2] Cai, Q., M. Brysbaert. 2010. SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE* 5.
[3] Chen, M. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.
[4] Chen, Y. 2012. Tonal Variation. In: A. C. Cohn, C. Fougeron, M. K. Huffman (eds), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 103-114.
[5] Fox, R. A., Y.-Y. Qi. 1990. Context Effect in the Perception of Lexical Tone. *J. of Chinese Linguistics* 18(2), 261-284.
[6] Francis, A. L., et al. 2003. On the (Non)Categorical Perception of Lexical Tones. *Perception & Psychophysics* 65(7), 1029-1044.
[7] Gandour, J. T. 1978. The Perception of Tone. In: V. A. Fromkin (ed), *Tone: A Linguistic Survey*. London: Academic Press, 41-76.
[8] Huettig, F., et al. 2011. Using the Visual World Paradigm to Study Language Processing: A Review and Critical Evaluation. *Acta Psychologica* 137, 151-171.
[9] Hyman, L. M. 2007. Universals of Tone Rules: 30 Years Later. In: T. Riad, C. Gussenhoven (eds), *Tones and Tunes*. Berlin: Mouton de Gruyter, 1-34.
[10] Li, Q., Y. Chen. 2012. Trisyllabic Tone Sandhi in Tianjin Mandarin. *Proc. 3rd International Symposium on Tonal Aspects of Languages* Nanjing, China.
[11] Li, Q., Y. Chen. under review. An Acoustic Study on Tone Sandhi in Tianjin Mandarin. *J. of Phonetics*.
[12] Li, X., S. Liu. 1985. Tianjin Fangyan de Liandu Biandiao (Tone Sandhi in Tianjin Mandarin). *Zhongguo Yuwen* 1, 76-80.
[13] Lin, H.-B., B. H. Repp. 1989. Cues to the Perception of Taiwanese Tones. *Haskins Laboratories Status Report on Speech Research* SR-99/100, 137-147.
[14] Malins, J. G., M. F. Joanisse. 2010. The roles of Tonal and Segmental Information in Mandarin Spoken Word Recognition: An Eyetracking Study. *J. of Memory and Language* 62, 407-420.
[15] Mirman, D. 2014. *Growth Curve Analysis and Visualization Using R*. Boca Raton: CRC Press.
[16] Moore, C. B., A. Jongman. 1997. Speaker Normalization in the Perception of Mandarin Chinese Tones. *J. Acoust. Soc. Am.* 102(3), 1864-1877.
[17] R Core Team (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing. http://www.r-project.org
[18] Shen, J., et al. 2013. On-line Perception of Mandarin Tones 2 and 3: Evidence from Eye Movements. *J. Acoust. Soc. Am.* 133(5), 3016-3029.
[19] Shen, X. S., M. Lin. 1991. A Perceptual Study of Mandarin Tones 2 and 3. *Language and Speech* 34(2), 145-156.
[20] Shen, X. S., et al. 1993. F0 Turning Point as an F0 Cue to Tonal Contrast: A Case Study of Mandarin Tones 2 and 3. *J. Acoust. Soc. Am.* 4(1), 2241-2243.
[21] Tanenhaus, M. K., et al. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science* 268, 1632-1634.
[22] Weber, A., et al. 2007. The Mapping of Phonetic Information to Lexical Representations in Spanish: Evidence from Eye Movements. *Proc. 14th ICPhS* Saarbrücken, 1941-1944.
[23] Xu, Y. 1994. Production and Perception of Coarticulated Tones. *J. Acoust. Soc. Am.* 95(4), 2240-2253.
[24] Xu, Y. 2001. Sources of Tonal Variations in Connected Speech. *J. of Chinese Linguistics* Monograph Series 17, 1-31.
[25] Zhang, J., J. Liu. 2011. Tone Sandhi and Tonal Coarticulation in Tianjin Chinese. *Phonetica* 68, 161-191.