Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/37609 holds various files of this Leiden University dissertation

Author: Qian Li
Title: The production and perception of tonal variation: evidence from Tianjin Mandarin
Issue Date: 2016-02-10
CHAPTER 4 TONAL VARIABILITY ON THE PERCEPTION OF LEXICAL TONES - EVIDENCE FROM EYE MOVEMENTS

4.1 Introduction

Among all languages in the world, 60-70% of them have lexical tones (Yip, 2002), where pitch changes over a word, together with other acoustic cues, signal lexical contrasts. In many tonal languages, tones are primarily signaled via different f0 patterns (Gandour, 1978, but see e.g., Andruski, 2006; Brunelle, 2009b for the effect of phonation types on tonal contrasts). A handful of studies have revealed that native speakers of tone languages utilize various types of f0 cues to identify lexical tones, such as the overall height of f0 contour (e.g., Lin & Repp, 1989; Shen et al., 1993; Francis et al., 2003; Shen et al., 2013), the direction of f0 change (e.g., Fox & Qi, 1990; Shen & Lin, 1991; Moore & Jongman, 1997), as well as the timing of f0 turning point (e.g., Shen & Lin, 1991; Shen et al., 1993; Moore & Jongman, 1997). These findings show the importance of f0 information in the perception of the canonical realization of lexical tones produced in isolation, but fall short in taking into consideration possible contextual f0 variation in connected speech, the perception of which on tonal identification is the most practiced behavior during daily speech communication.

When lexical tones are produced in connected speech, the f0 realization of the tones may deviate greatly from that produced in isolation, due to contextual effects such as tone sandhi and tonal coarticulation. Tone sandhi refers to phonological tonal alternation, which usually causes such a great change in the f0 contour that the sandhi-derived tones are typically unpredictable or unexpected from the canonical f0 realization. For instance in Beijing Mandarin, when two Low tones are combined in a disyllabic domain, the first Low tone, which otherwise would be realized with a low f0 target, would surface with a rising f0 contour (Chen, 2000 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 much subtler and with varying degrees of phonetic modification depending upon the tonal contexts. They are nevertheless rather predictable from their canonical f0 realization despite the phonetic modifications. This is known as tonal coarticulation (see Xu, 2001; Chen, 2012 and references therein for further details on tonal variation).

Despite the extensive contextual variation in the f0 realization of lexical tones in connected speech and the rich layers of different types of tonal variability, only a small number of perception studies have taken the contextual effect into consideration. This body of work shows consistently that listeners’ identification of a lexical tone is affected by its tonal context (e.g., Fox & Qi, 1990; Xu, 1994; Moore & Jongman, 1997; Francis et al, 2003). For example, Xu (1994) investigated the perception of the middle tone within

4 A version of this chapter has been submitted for publication as: Qian Li & Yiya Chen. (submitted). Tonal variability on the perception of lexical tones - evidence from eye movements. The Journal of the Acoustical Society of America.
trisyllabic constituents. The target was embedded in both “compatible contexts” in which its $f_0$ realization is less deviated, and “conflicting contexts” where its $f_0$ realization is greatly deviated. Results show that the listeners’ identification of lexical tones produced in both contexts is highly accurate with their respective original contexts (i.e., target produced in compatible contexts was presented in compatible contexts, and target produced in conflicting contexts presented in conflicting contexts). However, if the original tonal contexts were replaced by white noise, only tones in compatible condition can be correctly identified above chance level, while the identification of tones in conflicting condition dropped below chance level. Furthermore, when the target was presented in swapped tonal contexts (i.e., target produced in compatible contexts presented in conflicting contexts, and that produced in conflicting contexts presented in compatible contexts), listeners showed clear tendencies of compensating the influence of the tonal contexts when identifying the target, e.g., a rising tone that was flattened by a conflicting context in production can be perceived as a falling tone when it was presented in a new compatible context. This study thus indicates that when tones are greatly deviated from its canonical $f_0$ realization, listeners need to rely on the contextual tonal information in order to identify the target lexical tones. The question that remains interesting is how the target tones are processed in advance of access to the tonal contexts during speech recognition.

A second, hitherto independent, line of studies on contextual tonal perception has mainly focused on the issue of whether the sandhi-derived $f_0$ realization of lexical tones is perceptually distinguishable from another lexical tone (which has been proposed to be the targeted sandhi tone in most literature on tone sandhi). For instance, a few studies have been done on the perception of the Low tone sandhi in Beijing Mandarin. Although there is evidence that the sandhi-derived Low tone variant is processed differently from the lexical Rising tone in production encoding (Chen et al., 2011; Zhang et al., 2014; Nixon, Chen, & Schiller, 2015), and also articulated with slight acoustic differences from the Rising tone (e.g., Yuan & Chen, 2014 and references therein), perception studies have repeatedly shown that the Low+Low tonal sequence cannot be reliably distinguished from that of the Rising+Low sequence (e.g., Wang & Li, 1967; Speer et al., 1989). However, the issue of tonal perception in sandhi context is not limited to whether the tone sandhi patterns are neutralized in perception. Moreover, the end-state responses are likely to exhibit a different pattern from that of the real-time processing data (Spivey, 2007; also see discussion in Malins & Joanisse, 2010). Therefore even if listeners cannot reliably distinguish two tonal sequences in traditional meta-linguistic perceptual judgement tasks, it is still possible that the two tonal sequences are actually processed differently in spoken word recognition. So, the question that remains to be addressed in the literature is how lexical tones are processed during spoken word recognition, due to their allophonic sandhi variation.

The present study aims to fill the knowledge gaps on tonal perception by addressing the above two research questions. Via a word-recognition task within the Visual World Paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), this study sets out to investigate the processing of tonal variability in Tianjin Mandarin, which presents a good testing case as it exhibits interesting patterns of tonal variability in connected speech.
4.1.1 Tianjin Mandarin

Tianjin Mandarin is a dialect of Mandarin which is spoken in the urban areas of Tianjin city (about 100km from Beijing). Like Standard Chinese, Tianjin Mandarin has four lexical tones: Tone 1 (T1) is a low-falling tone, Tone 2 (T2) a high-rising tone, Tone 3 (T3) a low-dipping tone, and Tone 4 (T4) a high-falling tone (Zhang & Liu, 2011; Li & Chen, 2012, 2016). In connected speech, Tianjin Mandarin shows a range of complex tone sandhi patterns over disyllabic constituents. The received wisdom has been that these sandhi patterns involve the categorical change of one lexical tone to another. For example, when two T3s are combined, the first T3 is claimed to change into T2 (e.g., Li & Liu, 1985; Chen, 2000; Hyman, 2007). However, recent experimental data have shown that there is no complete tonal neutralization in Tianjin Mandarin, which means there is no categorical change of one lexical tone to another (Zhang & Liu, 2011; Li & Chen, 2012, 2016). Furthermore, two different types of tone sandhi have been observed in this language, one is Near-Merger Sandhi and the other No-Merger Sandhi, as illustrated in Figure 4.1.

![Figure 4.1](image_url)

**Figure 4.1** f0 realization of three disyllabic sandhi sequences in Tianjin Mandarin (T3T3 in a, T1T1 in b and T4T1 in c) compared to their respective sandhi target sequences (T2T3 in a, T3T1 in b, T2T1 in c) as claimed in the literature. T3T3 is Near-Merger Sandhi; T1T1 and T4T1 are No-Merger Sandhi. Thick white lines indicate the mean f0 of the sandhi sequences (dark gray areas for ±1 standard error of mean); black lines indicate the mean f0 of the claimed target sequences (light gray areas for ±1 standard error of mean). Normalized time.

In the Near-Merger Sandhi case, i.e., the T3T3 sequence in Figure 4.1a, the first T3 is realized with a high-rising f0 contour, which makes the f0 realization of this tonal sequence hardly distinguishable from that of the claimed target sequence (T2T3) in the literature. Figure 4.1a compares the f0 contour of T3T3 to that of T2T3. The data were averaged across 72 tokens (six speakers, three items, two informational status, two repetitions) (Li & Chen, 2012, 2016). White lines with dark areas represent the f0 realization of the sandhi pattern T3T3; and black lines with light areas represent the claimed target sequence T2T3. It can be seen from Figure 4.1a that, the f0 realization of the two sequences are hardly distinguishable despite some subtle differences.

In the No-Merger Sandhi cases, 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 Tianjin Mandarin: T1T1 (Figure 4.1b) and T4T1 (Figure 4.1c).
T1T1 has been claimed to change into T3T1 in the literature (e.g., Li & Liu, 1985; Chen, 2000; Hyman, 2007). As shown in Figure 4.1b, the first T1 in T1T1 is realized with a slightly falling and rising \( f_0 \) tone, which is clearly unexpected from its canonical low-falling \( f_0 \) contour when produced in isolation. Although T1\text{\textsubscript{sandhi}} is realized similarly to T3 as in T3T1, its \( f_0 \) realization is significantly different from that of T3, suggesting no merging of T1\text{\textsubscript{sandhi}} with T3. Similarly, T4T1 is claimed to change into T2T1 (e.g., Li & Liu, 1985; Chen, 2000; Hyman, 2007). While the first T4\text{\textsubscript{sandhi}} in T4T1 is similar to the T2 \( f_0 \) contour, they are clearly different from each other (Figure 4.1c).

In short, the two types of tone sandhi differ in how closely the sandhi-derived \( f_0 \) contour resembles one of the lexical tones in the lexical tone system. In the Near-Merger case, the sandhi-derived tone is realized with a \( f_0 \) contour that is hardly distinguishable from its target lexical tone, while in the No-Merger cases, the sandhi-derived lexical tones show a similar \( f_0 \) contour to their corresponding target lexical tones while still maintaining their distinctiveness. 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?

4.1.2 Visual world paradigm and spoken word recognition

An auditory word-recognition task within the Visual World Paradigm (VWP) (Tanenhaus et al., 1995) was employed to tap into the processing of lexical tonal variation. A typical setup of the VWP in speech recognition studies is to present participants with four pictures or written words on a computer screen. The participants simultaneously hear an auditory stimulus, which corresponds to one of the four possibilities (i.e., the target) while their eye movement is tracked (see a review in Huettig, Rommers, & Meyer, 2011). Typically, a target is presented with a competitor and two distractors. The participants were asked to identify the words they have just heard and click on the target with a mouse.

Based on the assumption that eye movements are closely time-locked to the spoken-word processing, this paradigm makes it possible to tap into the time-course of auditory word recognition (e.g., Tanenhaus et al., 1995; Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Beddor, McGowan, Boland, Coetzee, & Brasher, 2013; see also reviews in Tanenhaus, Magnuson, Dahan, & Chambers, 2000 and Huettig et al., 2011). Furthermore, the task involved in this paradigm is more natural speech perception experience than just asking the participants to make meta-linguistic judgements, which is typically used in traditional studies on tonal perception.

To make inferences about how the target auditory stimuli are processed, previous studies have mainly looked at how fast listeners start to fixate upon the target words (or competitors) as well as how long they look at the targets (or competitors). For example, in Tanenhaus et al. (1995), participants were asked to follow the instructions, e.g., “pick up the candy”, and move real objects around. When the target object (e.g., candy) was presented together with another object (e.g., candle) whose name has phonological overlapping with the target, it took the participants longer time (i.e., 230ms) to initiate an eye movement to
the target object candy than when a phonologically similar object was absent (i.e., 145ms). This demonstrates that the speech recognition process actually starts before participants hearing the end of the word. In Allopenna et al. (1998), participants were presented with displays of four images on the computer screen, one for target corresponding to the auditory stimuli (e.g., beaker), one for “onset competitor” which has onset overlapping with the target (e.g., beetle, which shares the onset with beaker), as well as two for distractors that are phonologically unrelated with the target (e.g., dolphin, carriage). The results show that when participants heard the instruction “pick up the beaker”, the proportion of looks to both beaker and beetle started to increase initially. However, as the speech unfolded, the proportion of looks to beettle started to decrease while only the picture corresponding to target gained more looks. In another experimental condition, the target words were presented with a “rhyme competitor” (e.g., speaker, which shares the rhyme with the target beaker). Compared to that of the “onset competitor”, the initial increase of looks to “rhyme competitor” speaker occurred at a much later time point since it only overlapped with the target at a later part of the word. This indicates that, upon hearing a spoken word, listeners continually evaluate the unfolding speech with fine-grained sensitivity, and gradually activate certain lexical candidates that compete for recognition.

Recently, VWP has been successfully employed in perception studies of lexical tones produced in isolation (Malins & Joanisse, 2010; Shen et al., 2013). For example, in Shen et al. (2013), participants were presented with images corresponding to segmentally identical T2 and T3 monosyllabic words (e.g., bi² ‘nose’ vs. bi³ ‘pen’) together with two other phonologically unrelated distractors (e.g., che¹ ‘car’, dao¹ ‘knife’). They were played with T2 or T3 stimuli and asked to click on the corresponding images. Stimuli were normalized with 500ms duration. Both T2 and T3 were resynthesized so as to have identical f0 contours from the onset until 200ms into the tones; from 200ms on, T2 was resynthesized with high f0 offsets and T3 low offsets. Results show that like segments, the recognition of lexical tones starts before the end of the stimuli and the processing of lexical tones is in a similar incremental fashion as that observed for segments. For example, upon hearing a T3 stimulus, participants generally needed only around 350-450ms to initiate an eye movement to the image corresponding to the target. Given that it usually takes about 200ms to plan and execute an eye movement (Matin, Shao, & Boff, 1993), it can be inferred that participants have already recognized the tone as early as around 150-250ms after the stimulus onset, which is much earlier than the end of the stimulus (500ms after stimulus onset). Furthermore, as the time unfolds, the proportion of looks to T3 continues to increase from 30% up to 90%, while that for other possibilities drops to the minimum at the same time.

This present study extended the VWP into studies of lexical tone perception in connected speech, where disyllabic collocations in Tianjin Mandarin with different types of tonal variability were used as auditory stimuli (i.e., Near-Merger Sandhi, No-Merger Sandhi, No Sandhi). The first goal of this study is to establish whether lexical tones produced in contexts are comparable with those produced in isolation in terms of the eye movement patterns during auditory speech recognition; second, it is also of great interest to investigate how different types of tonal variability affect online tonal processing. Note that the present
study is not particularly interested in whether certain sandhi sequence is ambiguous with its target sequence as claimed in the literature with segmental overlap for both syllables, e.g., \( zhi^3 fa^3 \) ‘fingering’ vs. \( zhi^3 fa^1 \) ‘executing justice’. Rather, the main focus is only on the perception of the first tone of different tonal variability, where there should be no ambiguity in the visual world context, e.g., \( zhi^3 fa^3 \) ‘fingering’ vs. \( zhi^4 hou^4 \) ‘lag’.

To address the first issue, two competitors for each target were included: a baseline competitor which does not share segments or tones with the target, and a segmental competitor which has overlapping segments but unrelated tones with the target (following the name in Malins & Joanisse, 2010 for the ease of comparison). Malins and Joanisse (2010) show that when targets and competitors are segmentally identical but different in tones (e.g., \( chuang^2 \) ‘bed’ vs. \( chuang^4 \) ‘window’), there is greater competition between targets and competitors than in the baseline condition with no segmental overlap. Similarly, greater competition effect for segmental competitors than for baseline competitors should also be expected in the present study.

To achieve the second goal, three types of tonal variability were included for the target words, Near-Merger Sandhi, No-Merger Sandhi, and No Sandhi. It can be predicted that 1) for Near-Merger Sandhi, the first tone is processed with the most effort due to the altered \( f0 \) realization and the potential ambiguity with another lexical tone in the tonal system; 2) No-Merger Sandhi should be less difficult to process than Near-Merger Sandhi, since its realization is altered but it is not supposed to be ambiguous with another tone in the tonal system; and 3) in the No Sandhi condition, the tonal realization of the first tone can be expected from the canonical realization, and thus no tonal ambiguity is involved; therefore, it should be processed with the least effort.

4.2 Method

4.2.1 Participants

Thirty-four native speakers of Tianjin Mandarin were recruited to participate in this experiment. All subjects were born in late 1980s or early 1990s (9 males, 25 females; Mean=22) and raised in the urban areas of Tianjin. They were undergraduate or postgraduate students studying in Beijing at the time of the experiment. None of them had lived out of the Tianjin city before 18. They were paid for their participation but unaware of the purpose of the experiment. One subject was wearing cosmetic contact lens, resulting in potentially different eye movement data. Her data were excluded from further analyses. Data from another two subjects who were shortsighted without wearing glasses for correction were also excluded. The remaining 31 subjects had normal or corrected-to-normal vision. All participants provided written informed consent.

4.2.2 Stimuli

The target stimuli consisted of 36 highly lexicalized disyllabic collocations with two sandhi patterns in Tianjin Mandarin: near-merger sandhi (i.e., T3T3) and no-merger sandhi (i.e.,
In addition, 18 stimuli with only tonal coarticulation and no sandhi changes were included for further comparison (i.e., T4T4, T3T2, T3T4) (see Table 4.1). Target stimuli of different tonal variability types were matched in bigram mutual information which stands for the likelihood of two syllables co-occurring within a lexical item according to Da (2004). The mean bigram co-occurring frequency of all target stimuli was 5.0 (SD=3.7), which corresponds to strong collocation strength, and the three groups of stimuli were not significantly different from each other (F(2)=0.26, p=0.773).

### Table 4.1 Experimental design and sample stimuli.

<table>
<thead>
<tr>
<th></th>
<th>Target</th>
<th>Baseline Competitor</th>
<th>Distractor 1</th>
<th>Distractor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Near-Merger Sandhi</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>endless</td>
<td>雨水 (yu³ shui³)</td>
<td>拖鞋 (tuo¹ xie²)</td>
<td>抽血 (chou¹ xie³)</td>
<td>犀牛 (xi¹ niu²)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slips</td>
<td>to draw blood</td>
<td>rhinoceros</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Segmental Competitor</td>
<td>Distractor 1</td>
<td>Distractor 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>預購 (yu⁴ gou⁴)</td>
<td>分号 (fen¹ hao⁴)</td>
<td>報警 (bao⁴ jing³)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>purchase in advance</td>
<td>semicolon</td>
<td>to report the police</td>
</tr>
<tr>
<td><strong>No-Merger Sandhi</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>浴缸 (yu⁴ gang¹)</td>
<td>冷气 (leng³ qi⁴)</td>
<td>消防 (xiao¹ fang²)</td>
<td>激動 (ji⁴ dong³)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cold air</td>
<td>fire-fighting</td>
<td>excited</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Segmental Competitor</td>
<td>Distractor 1</td>
<td>Distractor 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>语录 (yu³ lu⁴)</td>
<td>沙拉 (sha¹ la¹)</td>
<td>冬天 (dong¹ tian¹)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>quotation</td>
<td>salad</td>
<td>winter</td>
</tr>
<tr>
<td><strong>No Sandhi</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>继续 (ji³ xu⁴)</td>
<td>缆车 (lan³ che¹)</td>
<td>雨衣 (yu³ yi¹)</td>
<td>手镯 (shou³ zhuo²)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cable car</td>
<td>raincoat</td>
<td>bracelet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Segmental Competitor</td>
<td>Distractor 1</td>
<td>Distractor 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>挤压 (ji³ ya¹)</td>
<td>椭圆 (tuo³ yuan²)</td>
<td>海港 (hai³ gang³)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to press</td>
<td>oval</td>
<td>harbor</td>
</tr>
</tbody>
</table>

For each target, a corresponding baseline competitor and a segmental competitor were chosen (a within-item design). The baseline competitors did not share segment or tone with the targets (S-, T-). The segmental competitors shared the segment of the first syllable with the targets, but with an unrelated tone (S+, T-), which was neither the underlying lexical tone nor the lexical tone which was claimed in the literature as the targeted sandhi tone. The second syllables of the targets and competitors were different in terms of both tone and segment across all conditions. Take the target "zhi³ fa³ ('fingering') as an example: its baseline competitor was "wai⁴ tao⁴ ('coat') while the segmental competitor
was “zhì hòu” (‘lag’). Each target-competitor pair also had two distractors within the Visual-World Paradigm. The distractors did not share any tone or segments with the target or the competitor.

The targets and competitors were further controlled to be closely matched in terms of lexical frequency based on Cai and Brysbaert (2010) and orthographic complexity as we presented Chinese characters instead of pictures (following Shen et al., 2013), so that there was no significant difference between the targets and competitors for lexical frequency (F(1)=0.087, p=0.768) or visual complexity (F(1)=0.156, p=0.694). No participant reported difficulty of recognizing any character.

The auditory stimuli were pre-recorded by a male speaker of Tianjin Mandarin who was born in the 1980s. All 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 was not significantly different across the three conditions (F(2)=0.47, p=0.627); the mean duration of the entire disyllabic stimuli was 704ms (SD=98ms).

4.2.3 Procedure

Eye movements were recorded with an Eyelink 1000 system with a 35mm lens running at 500Hz. Visual stimuli were presented on a 21-inch ViewSonic G220f monitor (resolution: 1024*768 pixels; frame rate: 120Hz). Participants were seated comfortably with a chin rest and a forehead rest set at a distance of 69cm from the screen. All recordings and calibrations were done monocularly based on the left eyes and viewing was binocular (except for one participant whose right eye was tracked for tracking failure of the left eye). Gaze position was calibrated with a 13-point grid. Prior to the presentation of each trial, there was a drift check.

The experiment began with a block of 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 (500ms) with only a fixation cross at the very center of the screen prior to the stimuli screen. The participants were asked to look at the fixation cross until it disappeared. Then the auditory stimulus was played through a headphone simultaneously with the presentation of the visual stimuli screen. The task for the participants was to click on the lexical item they had heard with a mouse within 10 seconds (although no participant spent more than 10 seconds to make a decision).

In each trial, participants saw four disyllabic collocations on the screen, which consisted of a target corresponding to the auditory stimulus, a competitor, and two distractors. To maximally guarantee the non-overlapping of the parafoveal view when reading each collocation (e.g., Mielle et al., 2009), size and location of the visual stimuli were calculated accordingly to the screen (2.5cm * 5cm in height and width, corresponding to 5° visual angle; approximately at the centers of the four quadrants of the screen; for further details, we refer readers to Appendix I). The order of the four visual stimuli on the screen was counter-balanced. The trial order was pseudo-randomized so that auditory stimuli of the same tonal variability condition were not presented in consecutive trials.
4.2.4 Eye movement data analysis

Given that the two characters of a disyllabic collocation on the screen basically covered the foveal area (see Figure I in Appendix I), only eye gaze data within the boundary immediately around the collocation were included for the analyses of respective stimulus item (i.e., target, competitor and distractors). The eye movement data were recorded at 2ms intervals. The proportion of looks at each time point (every 2ms) was calculated. Trials in which subjects clicked on items other than the target (<1%) were excluded from the analyses. For better illustration of the eye movement patterns, data from trials containing blinks (17%) were also excluded.

The eye movement data were reported from the onset of the auditory stimulus to 1400ms post stimulus onset. The upper limit was chosen at the point where the proportion of looks to target had reached the maximum (following Malins & Joanisse, 2010).

The growth curve analysis was used for statistics (following Mirman, 2014) with the package lme4 (Bates et al., 2014) in R (R Core Team, 2014). Given the large dataset obtained for this study, the statistical analysis was limited within certain time windows (i.e., part of the curves), which showed interesting patterns, instead of fitting the entire curves.

The base model was first constructed with only TIME (up to the fourth order components) in the fixed-factor structure and SUBJECT as the random factor on all time terms (following Mirman, 2010). Additional fixed factors were then assessed in a stepwise fashion. Only factors that significantly improved the model fits were added into the model. Parameter-specific \( p \)-values were estimated using the normal approximation (treating the \( t \)-value as a \( z \)-value).

4.3 Results

4.3.1 Baseline comparison

Figure 4.2 illustrates the mean proportion of looks to target averaged across three types of tonal variability, when the competitor is a segmental competitor (i.e., with segmental overlap) vs. when there is a baseline competitor (i.e., without segmental overlap). X-axis stands for the time since target onset and y-axis for the proportion of looks to target.

As can be seen from Figure 4.2, when there is segment overlap between the targets and the competitors (shown with solid line as SEGMENTAL), the proportion of looks to target shows a different pattern from that in the baseline competitor condition (shown with dotted line as BASELINE) mainly in two aspects. First, the proportion of looks to target in the segmental condition starts to increase at a much later time point (around 500ms) than that in the baseline condition (around 200ms). Second, the segmental condition shows lower overall proportion of looks to target than in the baseline condition, especially salient in the time window of 200-800ms, which approximately corresponds to about 600ms into the auditory stimuli (the first syllable is 357ms on the average). After 800ms, the differences between the two conditions are less dramatic.
A growth curve analysis was run for the comparison between segmental vs. baseline conditions within the 200-800ms time window. The full model included the interaction of the fourth-order time terms, COMPETITOR TYPE (two levels: SEGMENTAL & BASELINE) and TONAL VARIABILITY TYPE (three levels: NEAR-MERGER SANDHI, NO-MERGER SANDHI and NO SANDHI) as the fixed factor and SUBJECT as a random factor on all time terms. We also included FREQUENCY, NUMBER OF STROKES, BIGRAM MUTUAL INFORMATION and STRUCTURE of the target stimuli as additional fixed factors in the full model. The results confirmed a significant main effect of COMPETITOR TYPE over the time terms up to the third-order within the 200-800ms time window (intercept: $\beta=0.15$, $t=33.68$, $p<0.001$; slope: $\beta=-0.18$, $t=-7.39$, $p<0.001$; quadratic: $\beta=-0.13$, $t=-5.34$, $p<0.001$; cubic: $\beta=0.09$, $t=3.86$, $p<0.001$). This indicates that the proportions of looks to the target in two conditions were significantly different in the overall mean, the direction of curve change, as well as the steepness of the curve change.

4.3.2 Different tonal variability types

To further investigate the eye movement patterns in three types of tonal variability, we included analyses for looks to both targets and competitors in this section.

4.3.2.1 Looks to the target

Figure 4.3 shows the proportion of looks to target of three different tonal variability types when target and competitor shared segments. X-axis stands for the time since target onset, y-axis for the proportion of looks to target.
Figure 4.3 The proportion of looks to target when targets and competitors have segmental overlap in three tonal variability types, aggregated across participants and items (thick solid line for Near-Merger Sandhi, thin solid line for No-Merger Sandhi, dotted line for No Sandhi).

We can see from Figure 4.3 that, due to different types of tonal variability in the targets, there are subtle differences in the proportion of looks to the target. In general, the proportion of looks to target in all tonal variability types remains at the bottom (less than 0.1) until around 500ms, when the proportion starts to rise. This indicates that the participants were not looking at the targets until 500ms across all tonal variability types. However, after 500ms, the proportion of looks to target in different tonal variability types began to diverge, especially in two time windows: 500-700ms and 800-1100ms.

In the 500-700ms time window (WINDOW 1, Figure 4.3), the No-Sandhi condition shows a slightly faster increase in the proportion of looks to target than in the other two sandhi conditions. A growth curve analysis was run for the comparison among three tonal variability types within this window. The full model included the interaction of the fourth-order time terms and TONAL VARIABILITY TYPE (three levels: NEAR-MERGER SANDHI, NO-MERGER SANDHI and NO SANDHI) as the fixed factor and SUBJECT as the random factor on all time terms. Additional fixed factors included FREQUENCY, NUMBER OF STROKE, STRUCTURE and BIGRAM MUTUAL INFORMATION of the target, as well as NUMBER OF STROKE of the competitor. Although the difference is subtle, the growth curve analysis results suggested that the proportion of looks to the target in No-Sandhi condition had a significantly faster increasing rate than the other two conditions, in terms of the quadratic components (quadratic for Near-Merger Sandhi vs. No Sandhi: $\beta=0.05$, $t=1.99$, $p<0.05$; No-Merger Sandhi vs. No Sandhi: $\beta=0.05$, $t=2.04$, $p=0.04$). Two sandhi conditions were not significantly different from each other in any aspect within this time window.

The second interesting time window is from 800ms to 1100ms (WINDOW 2, Figure 4.3). Within this window, our full model of growth curve analysis included the interaction of the fourth-order time terms and TONAL VARIABILITY TYPE (three levels: NEAR-MERGER SANDHI, NO-MERGER SANDHI and NO SANDHI) as the fixed factor and SUBJECT as the random factor on all time terms. Additional fixed factors included
FREQUENCY, STRUCTURE and BIGRAM MUTUAL INFORMATION of the target stimuli, as well as STRUCTURE of the competitor. In this window, there are noticeable differences in the three conditions, among which two observations can be made.

First, No-Sandhi condition differs from the two sandhi conditions in terms of both the overall mean and the slope of proportion of looks at the target. There is clearly less proportion of looks at the target in the No-Sandhi condition, as suggested by the significantly results in the intercepts for Near-Merger Sandhi vs. No-Sandhi condition (intercept: $\beta=0.05$, $t=5.75$, $p<0.001$) and No-Merger Sandhi vs. No-Sandhi (intercept: $\beta=0.08$, $t=8.50$, $p<0.001$). In addition, the No-Sandhi condition had a significantly larger rising slope of the proportion of looks to the target than both Near-Merger Sandhi (slope for Near-Merger Sandhi vs. No-Sandhi: $\beta=-0.10$, $t=-3.12$, $p<0.01$) and No-Merger Sandhi (slope for No-Merger Sandhi vs. No-Sandhi: $\beta=-0.10$, $t=-3.29$, $p<0.001$). The slopes of two sandhi conditions did not significantly differ from each other.

Second, the two sandhi conditions showed significantly different proportions of looks to the target in terms of the overall mean. There are overall less looks to the target in Near-Merger Sandhi than that in the No-Merger Sandhi condition, as reflected in the significant results in the intercept between Near-Merger Sandhi vs. No-Merger Sandhi ($\beta=0.03$, $t=3.16$, $p<0.001$).

4.3.2.2 Looks to the competitor

![Figure 4.4](image-url)  
**Figure 4.4** The proportion of looks to competitor when target and competitor have segment overlap in three tonal variability types, aggregated across participants and items (thick solid line for Near-Merger sandhi, thin solid line for No-Merger sandhi, dotted line for No Sandhi).

If not at the target, where else could the listeners be looking? We further analyzed the eye movement patterns over competitor. Figure 4.4 illustrates the proportion of looks to competitor in three tone sandhi types. X-axis stands for the time since target onset, y-axis for the proportion of looks to competitor. The time frames in WINDOWS 2 and 3 in Figure 4.4 correspond to that of WINDOW 1 and WINDOW 2 in Figure 4.3, respectively.

In Figure 4.4, the proportion of looks to competitor in all tonal variability types
remains at the bottom, starting to rise and diverges after 200ms. Three time windows are particularly interesting: 300-400ms (WINDOW 1, Figure 4.4), 500-700ms (WINDOW 2, Figure 4.4) and 800-1100ms (WINDOW 3, Figure 4.4).

In the window of 300-400ms (WINDOW 1, Figure 4.4), the No-Sandhi condition shows relatively higher proportion of looks to the competitor than two sandhi conditions with a slight magnitude of difference. Taking a look at the whole curve for No-Sandhi condition in this figure, we could also see that the peak value for the No-Sandhi condition appears within the 300-400ms window. Growth curve analyses confirmed a significant difference in the overall mean between the Near-Merger Sandhi vs. No-Sandhi conditions (intercept: β=-0.07, t=-5.68, p<0.001) and between the No-Merger Sandhi vs. No-Sandhi conditions (intercept: β=-0.06, t=-5.04, p<0.001). Two sandhi conditions were not significantly different from each other.

In the time window from 500ms to 700ms (WINDOW 2, Figure 4.4), the proportion of looks to the competitor in the No-Merger Sandhi condition has a relatively higher mean than that in the other two conditions. This was confirmed by the significant difference in the overall mean between No-Merger Sandhi vs. No-Sandhi conditions (intercept: β=0.04, t=4.57, p<0.001) and No-Merger Sandhi vs. Near-Merger Sandhi conditions (intercept: β=0.03, t=3.13, p<0.01).

In the window from 800ms to 1100ms (WINDOW 3, Figure 4.4), the Near-Merger Sandhi condition shows relatively more overall proportions of looks to the competitor than both No-Merger Sandhi and No-Sandhi conditions. This could be confirmed by the significant difference in the intercept of the No-Merger Sandhi condition (intercept: β=-0.03, t=-5.27, p<0.001) and No-Sandhi condition (intercept: β=-0.03, t=-4.99, p<0.001), while No-Merger Sandhi and No-Sandhi were not significantly different from each other.

4.4 Discussion & conclusion

While it is well established that there is extensive variability in the f0 realization of lexical tones due to the neighboring tones, it is less understood in how native listeners perceive contextual tonal variations in spoken word recognition. The present study set out to shed some light on this issue with evidence from Tianjin Mandarin, which exhibits rich layers of contextual tonal variability.

We employed the Visual World Paradigm to investigate the time course of disyllabic tone perception with evidence from Tianjin Mandarin. We examined the participants’ looks to both targets and competitors within the Visual World Paradigm when they heard disyllabic stimuli with different types of tonal variability (i.e., Near-Merger Sandhi, No-Merger Sandhi, No Sandhi). Our results yield significant perceptual differences among different types of tone variation, which affects online speech processing differently, as reflected in the different eye movement patterns.

To first ensure the participants were behaving in a comparable way with what have been reported in studies of lexical tones produced in isolation, we included the baseline comparison where the targets and competitors had neither segmental nor tonal overlap. We found that when targets and competitors share the same segment for the first syllable,
there was a generally smaller proportion of looks to the target than that in the baseline condition. This indicates a stronger overall competition between targets and competitors when there is segmental overlap, which is comparable with what have been reported in a recent perception study on monosyllabic Mandarin tones (Malins & Joanisse, 2010). What differs between ours and the earlier study is that we found much earlier divergence of proportion of looks to target: only around 200ms post target onset in ours, but around 600ms in Malins and Joanisse (2010). This is probably due to the fact that we used printed words as the visual stimuli rather than images as used in Malins and Joanisse (2010). Printed words have been argued to be more sensitive to phonological manipulations than images, at least for alphabetic languages (e.g., Weber, Melinger, & Tapia, 2007; Huetig et al., 2011). Shen et al. (2013) compared image vs. character stimuli, but claimed no difference. However, their experimental setups for two types of stimuli were not completely comparable for comparison. Future studies are therefore needed to verify the potentially different effects of two types of visual stimuli on eye movement patterns in logographic languages during online speech recognition.

To look into how different kinds of contextual tonal variability affect tonal perception, we compared the proportions of looks to both targets and competitors in three different tonal variability conditions. As predicted, we have observed significant differences among the three different types of tonal variability. First, we have observed a different eye movement pattern upon hearing target stimuli with No-Sandhi tonal coarticulation from those in two sandhi conditions, with evidence from the gaze patterns over both targets and competitors. In terms of the proportion of looks to the target, No-Sandhi condition showed quicker increase of looks to the target compared to the two sandhi conditions in the time frame from 500ms to 700ms (WINDOW 1, Figure 4.3). Later on, participants did not seem to look at the No-Sandhi targets any more than they did in the two sandhi conditions in the time window from 800ms to 1100ms (WINDOW 2, Figure 4.3). Data of proportion of looks to the competitor suggested that the participants have shifted their attention from both targets and competitors to somewhere else at this time point, as in the time window from 800ms to 1100ms, the proportion of looks to the competitor in No-Sandhi condition was also lower than sandhi conditions (WINDOW 3, Figure 4.4). This, together with the finding that there are significantly more looks to competitors in the No-Sandhi condition than the two sandhi conditions at an earlier time point (300-400ms) (WINDOW 1, Figure 4.4), confirmed the relatively less effort required in processing No-Sandhi coarticulation, as the tones are realized with only subtle phonetic deviation.

Second, compared to the No-Sandhi coarticulation condition, two sandhi conditions have presented more challenge to listeners in tonal recognition due to the more dramatic f0 contour distortion. This was reflected in later initiation of fixation on the targets in both sandhi conditions (WINDOW 1, Figure 4.3). Furthermore, two sandhi conditions showed different eye movement patterns from each other for both targets and competitors. Between two sandhi conditions, No-Merger Sandhi seemed to be relatively easier to identify, as confirmed by the overall more looks to the target than Near-Merger Sandhi in the time window 800ms to 1100ms (WINDOW 2, Figure 4.3). The analyses of competitors
in Figure 4.4 further suggested that No-Merger Sandhi already showed more proportion of looks to the competitors than that of Near-Merger Sandhi at an earlier time point from 500ms to 700ms (WINDOW 2, Figure 4.4), while in the time window from 800ms to 1100ms (WINDOW 3, Figure 4.4), looks to competitors for No-Merger Sandhi significantly declined to the same low proportion as for the No-Sandhi condition. This suggested the successful recognition of No-Merger Sandhi stimuli at this time point, as participants have shifted their attention away from both target and competitor.

Compared to No-Merger Sandhi, Near-Merger Sandhi seems to be more difficult to process, reflected in significantly less looks at the target in the 800-1100 time window than No-Merger Sandhi. The significantly higher proportion of looks to the competitor for Near-Merger Sandhi in WINDOW 3 of Figure 4.4 further suggested that, within the time window from 800ms to 1100ms, the Near-Merger Sandhi stimuli have not been completely recognized as participants were still searching for the target.

Previous studies have shown clear effect of tonal contexts on lexical tone identification (e.g., Lin & Wang, 1985; Fox & Qi, 1990; Moore & Jongman, 1997; Francis et al., 2003), especially when tones largely deviate from its canonical realization, where listeners need to rely on the contexts to identify the tones (Xu, 1994). On the one hand, our study has lend further support to this contextual effect; on the other hand, by extending this body of literature, 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.

For the No-Sandhi condition, where there is only minor variation while tonal distinctiveness is maintained, listeners relied relatively less on the contexts in recognizing the tone of the first syllable as reflected in the relatively faster growth in the proportion of looks to the target. For two sandhi conditions, the $f_0$ realization of the lexical tones was altered to such a great extent that the tonal distinctiveness was no longer kept. It would be difficult to identify the tone only based on the sandhi-derived $f_0$ realization (also as suggested in Zhou & Marslen-Wilson, 1997). Information of the second syllable was thus needed to a greater extent to identify the lexical tone of the first syllable, which was reflected in the much later looks to target in both conditions.

This thus has provided clear perceptual evidence in support of the distinction between two contextual tonal variation processes, i.e., tone sandhi vs. tonal coarticulation. Although there is some consensus on different $f_0$ variation triggered by tone sandhi vs. tonal coarticulation, previous impressionistic studies claim no essential distinction between the two processes (e.g., Chen, 2000). Our data showed that the different degrees of $f_0$ deviation due to tone sandhi and tonal coarticulation indeed have different perceptual consequences in the process of online speech recognition. This thus suggests the necessity to tease apart the two processes with care when one attempts to analyze and account for the phenomenon of contextual tonal variability.

Our results further suggested the need to differentiate two types of tone sandhi in Tianjin Mandarin. These different types of tonal variability over the first syllable clearly have different consequences on tonal perception; because otherwise, different eye movement patterns for different tone sandhi types could not have been observed. For No-
Merger Sandhi where sandhi tones are rather distinct from the canonical tonal realization of any lexical tone within the lexicon, the difficulty in identification is mainly due to the recognition of a distorted $f_0$. Near-Merger Sandhi, comparatively, is more difficult to identify, given that the sandhi-derived tones are indeed quite undistinguishable from another lexical tone in the system. The successful identification thus involves both recognition of a distorted $f_0$ and the competition between two lexical tone candidates, which therefore requires more effort than for No-Merger Sandhi cases. While traditional perceptual studies on contextual tone sandhi variation mainly rely on native listeners’ metalinguistic knowledge of whether one tone is changed into another due to tone sandhi (e.g., Wang & Li, 1967; Speer et al., 1989), our paper is the first to look into how lexical tones with different types of tone sandhi are processed online.

Taking together, we have demonstrated that despite the commonly accepted “categorical perception” of lexical tones (e.g., Chan, Chuang, & Wang, 1975; Burnham & Jones, 2002; Francis et al., 2003; Xu, Gandour, & Francis, 2006; Xi, Zhang, Shu, Zhang, & Li, 2010; Zhang, Xi, Wu, Shu, & Li, 2012; but see Abramson, 1977 for non-categorical perception in Thai), native listeners are quite sensitive to fine-grained $f_0$ details in tone perception, which is in line with Malins and Joanisse (2010) and Shen et al. (2013). We observed a graded time course difference in online perception of contextual tonal variants, depending on how much the variants deviate from the canonical $f_0$ realization and whether there is competition with another lexical tone within the tonal inventory. Specifically, No-Sandhi tonal coarticulation was the easiest to recognize as the proportion of participants’ looks to the target increased with the fastest rate among three conditions; Near-Merger Sandhi was more difficult to process than No-Merger sandhi, reflected in the slower increase of looks to the target as well as in the overall less proportion of looks to target. The further implications of our findings on possible linguistic theory of tonal variability and sandhi alternation needs to be explored in the future.