Identification of acoustically modified Mandarin tones by native listeners

Chao-Yang Lee\textsuperscript{a,}\textsuperscript{*}, Liang Tao\textsuperscript{b}, Z.S. Bond\textsuperscript{b}

\textsuperscript{a}School of Hearing, Speech and Language Sciences, Grover Center W225, Ohio University, Athens, OH 45701, USA
\textsuperscript{b}Department of Linguistics, Ohio University, Athens, OH 45701, USA

Received 7 February 2007; received in revised form 15 January 2008; accepted 22 January 2008

Abstract

This study investigated Mandarin tone identification by 40 native listeners when only partial acoustic information was available. Twelve minimal tone pairs including all six Mandarin tonal contrasts were digitally processed to generate four syllable modifications: intact, silent-center, center-only, and onset-only. The syllables were recorded in two carrier phrases such that the offset of the carrier tone and the onset of the test tone were either a match or mismatch in fundamental frequency ($f_0$). In three experiments, the test syllables were presented in the original carrier phrases, excised from the carrier phrases, or excised and cross-spliced with another carrier phrase. Response accuracy and reaction time (RT) were measured. Listeners identified the tones at better than 86% correct with or without the carriers except when they heard the onset-only syllables in isolation, when their identification accuracy fell to 73% but still beyond chance. The modifications impacted the four tones differently, with Tones 1 and 2 being compromised more than Tones 3 and 4. Confusion matrix analyses showed that Tone 2 was predominantly confused with Tone 3, and Tone 1 was primarily confused with Tone 4. There were no main effects of splicing or match/mismatch with the carrier tone; however, in the cross-spliced context, syllables originally produced with a matching carrier tone were identified faster and/or more accurately. The implications of these findings for lexical tone processing are discussed.

Published by Elsevier Ltd.

1. Introduction

Speech perception involves extracting information from the acoustic signal to uncover the linguistic representations in a spoken message. Despite the many sources of variability in the acoustic input, human listeners are adept in making use of multiple acoustic cues and phonetic knowledge to achieve spoken language comprehension. The perception of impoverished acoustic signals has been used to identify the acoustic cues to phonetic distinctions and the phonetic knowledge involved in speech perception. For example, Strange and her colleagues examined vowel perception in “silent-center” consonant-vowel (CV) syllables (Jenkins, Strange, & Edman, 1983; Strange, 1989; Strange, Jenkins, & Johnson, 1983). These syllables were constructed by silencing the majority of the syllable center, rendering the steady-state formant frequency information unavailable.

\textsuperscript{*}Corresponding author. Tel.: +1 740 593 0232; fax: +1 740 593 0287.
\textit{E-mail address:} leec1@ohio.edu (C.-Y. Lee).

0095-4470/$ - see front matter Published by Elsevier Ltd.
Despite the absence of the static formant frequency information, which was traditionally considered the primary acoustic correlate to vocalic distinctions, listeners remained highly accurate in identifying vowels in these syllables. This result indicates that dynamic spectral information from CV transitions was useful in uncovering vowel identity.

The research reported here examined Mandarin tone identification from limited acoustic input. In particular, this study attempted to extend Gottfried and Suiter (1997), who applied the silent-center paradigm to investigating the perception of Mandarin vowels and tones. In Mandarin, tones are lexically contrastive just as segmental phonemes are. The primary acoustic correlate to Mandarin tones is the fundamental frequency (f0) pattern over a syllable. When produced in isolation, the four tones in Mandarin are generally described as high-level (Tone 1), rising (Tone 2), low (Tone 3), and falling (Tone 4) (e.g., Shih, 1987). Gottfried and Suiter (1997) constructed four types of test CV syllables: intact, center-only (initial six and final eight pitch periods were silenced), silent-center (all but the initial six and final eight periods were silenced), and onset-only (all but the initial six periods were silenced). The test syllables were presented in isolation or with a following syllable zi (字, meaning “word”). In both conditions, native listeners made few identification errors except for the initial-only syllables. Although the majority of the syllable was missing in the silent-center syllables, the identification of these syllables was as accurate as the intact and center-only syllables. This pattern of results is similar to what was found for vowel identification in English (e.g., Strange et al., 1983), suggesting that Mandarin listeners were also able to integrate information from both the initial and the final portions of a syllable to uncover the vowels and tones.

Tone identification from limited acoustic input has also been investigated in several other studies. Although the syllable is generally considered the tone-bearing unit for Mandarin, findings from these perception studies indicate that native listeners do not need the entire syllable to identify tones reliably. Tseng (1981) presented fragments of isolated vowels to native Mandarin listeners and found reliable tone identification with approximately half of a vowel. Similarly, Yang (1992) showed that isolated Mandarin tones could be correctly identified when half of the voiced part of a syllable was present. More recently, Liu and Samuel (2004) presented isolated Mandarin syllables with f0 information either partially masked by noise or completely removed by using “whispered” speech. Despite the missing f0 information, tone identification remained fairly accurate, suggesting that listeners were able to make flexible use of other acoustic cues when f0 information was compromised or unavailable.

Compared with these studies of tone perception in isolated syllables, the results from Gottfried and Suiter (1997) further implicated the role of tonal context. In particular, they found that the pattern of some tone confusions changed substantially depending on the presence of a tonal context. For example, when the following syllable zi was present, Tone 4 of the initial-only test syllable was only occasionally misidentified as Tone 1 (13%). In contrast, when the context was removed, the confusion rose to 66%. Their acoustic analyses showed that the context zi had a higher f0 onset after Tone 1 and a lower f0 onset after Tone 4. Gottfried and Suiter (1997) interpreted this result as listener’s use of the low f0 onset in the context zi as evidence for Tone 4 in the test syllable. When the context was removed, listeners could no longer use the contextual information; rather, they had to rely only on the initial six periods in the test syllable for tone identification. Since the f0 of the first six periods for Tone 4 was similar to that for Tone 1, the Tone 4→Tone 1 confusion became more prevalent.

In fact, contextual tonal variation has been noted extensively in the literature. Acoustic studies have shown that the canonical f0 contour of a tone is modified by neighboring tones in connected speech (Abramson, 1979; Han & Kim, 1974; Shen, 1990; Shih, 1987, 2005; Shih & Sproat, 1992; Wu, 1984, 1988; Xu, 1994, 1997). The degree of the modification appears to vary depending on the nature of the context. For example, examining trisyllabic, non-word sequences in Mandarin, Shen (1990) found no effect of adjacent tones on the f0 direction of the middle tone when speakers were instructed to read the syllables with evenly distributed stress. Other studies, in contrast, have found that the modification could be so substantial that the resulting f0 contour failed to resemble its canonical f0 contour. Using three- and four-syllable, real-word Mandarin sequences, Shih and Sproat (1992) and Shih (2005) showed that the f0 contour of the rising tone in a high-level tone context approached that of the high-level tone when spoken in a prosodically weak position. Using trisyllabic, real words or phrases in Mandarin, Xu (1994) found that the rising tone in the middle of the three-syllable sequences often became falling when the tonal context had conflicting f0 values with the middle tone. It
appears that how the stimuli were produced and perhaps the lexicality of the stimulus items contributed to the difference.

Compared with these acoustic studies on contextual tonal variation, there are less perceptual data on the effect of tonal coarticulation. Xu (1994) showed that the deviation from the canonical $f_0$ contour did not appear to be a problem for native listeners. In that study, naturally produced Mandarin trisyllabic words or phrases were processed such that all semantic information was removed but the tonal information remained. The test tone—the second tone in a three-tone sequence—was presented in the original context, in isolation, or in a “swapped” context where the first and third tone in the sequence was swapped. When presented in the original context, tone identification accuracy was high whether the tonal context was conflicting with the test tone or not. When presented in isolation, identification was accurate only when the context was not conflicting. When presented in the swapped context, tone identification was heavily influenced by the swapped context in addition to the $f_0$ contour of the test tone itself. These findings demonstrated that listeners do take into consideration contextual tonal variation in tone perception.

Reliable tone identification from limited acoustic input and the potential use of tonal coarticulatory information in tone perception was also shown in a gating study by Lee (2000). The gating task (Grosjean, 1996) is commonly used in studies of spoken word recognition to examine the precise amount of acoustic information needed for word recognition. In Lee’s (2000) experiment, successive fragments of Mandarin monosyllabic words, embedded in a semantically neutral carrier phrase, were presented to native Mandarin listeners for word recognition. The result showed that the listeners were able to reliably identify whether the word had a high $f_0$ onset tone (Tones 1 or 4) or a low $f_0$ onset tone (Tones 2 or 3) with as little as 20–40 ms of $f_0$ information into the voiced portion of the test syllable. Compared with earlier investigations showing that tone identification required half of an isolated syllable (Tseng, 1981; Yang, 1992), Lee’s (2000) finding suggests that listeners were able to derive tone identity not only from the vowel or syllable rhyme but also from the transition of a preceding tone to the underlying pitch targets of the test syllable (Xu, 2004; Xu & Wang, 2001). This result is therefore consistent with the idea that tonal coarticulation is a useful source of information for tone perception (Xu, 1994).

Compared with Xu’s (1994) approach of removing all semantic information in the acoustic signal while keeping tonal information available, Gottfried and Suiter (1997) hinted that tonal coarticulatory information could be useful even when the acoustic input was brief and incomplete. In contrast to the perceptual studies on isolated tones (Liu & Samuel, 2004; Tseng, 1981; Yang, 1992), the use of a following tone in the stimuli allowed Gottfried and Suiter (1997) to evaluate the effect of the presence/absence of a tonal context. What remains unclear is whether a preceding context could also provide useful information for tone identification from brief acoustic input. From the point of view of real-time speech processing, the role of the preceding context may be of equal, if not greater, importance. Speech is temporal. There is also ample evidence from the speech perception and spoken word recognition literature that listeners make use of information as soon as it becomes available in the acoustic signal (e.g., Cutler & Chen, 1997; Marslen-Wilson, 1987). Furthermore, although the bidirectional nature of contextual tonal variation has been established in the acoustic studies, there are conflicting findings on the magnitude and direction of the contextual effect. Shen (1990) found the effect of a preceding tone (carryover) to be similar to that of a following tone (anticipatory). Other studies, however, suggest that the anticipatory and carryover effects were asymmetric (Shih & Sproat, 1992; Xu, 1994, 1997). While Gottfried and Suiter (1997) implicated the use of carryover information in tone perception, the anticipatory information could be useful as well. Lee’s (2000) gating experiment was the only study other than Gottfried and Suiter (1997) to examine tone identification from brief acoustic input. Although the findings did suggest that listeners exploit anticipatory coarticulatory information for tone identification, no isolation condition was included in that study to allow a direct comparison with the context condition.

The purpose of this research was to further evaluate how well native listeners identify Mandarin tones from limited acoustic input and the role of tonal context in tone identification. Naturally produced Mandarin syllables were digitally processed to generate intact, center-only, silent-center, and onset-only syllables, which were then presented to native listeners for tone identification. Given the scarcity of perceptual data on tone identification from partial acoustic input, the present study adopted the same signal processing strategy as that used in Gottfried and Suiter (1997) to allow a direct comparison. Furthermore, the present study extended Gottfried and Suiter (1997) in several ways. First, tone identification was evaluated in three different contexts:
in isolation, in the original carrier phrase, and in a “cross-spliced” context. In particular, each test syllable was recorded in two carrier phrases. The isolation condition stripped away the tonal context, and the cross-spliced context presented potentially conflicting $f_0$ information, similar to Xu’s (1994) “swapping” context. Second, the offset $f_0$ of the carrier tone immediately preceding the test tone was chosen to be either a match or mismatch with the onset $f_0$ of the test tone. Xu’s (1994) acoustic analyses showed greater deviation of a tone from the canonical $f_0$ contour when its $f_0$ value mismatches that of a neighboring tone. His perceptual data also showed that tone identification in the mismatching context remained accurate only with the original carrier, but not when the carrier was altered or removed. The two context manipulations introduced in the present study should also provide information on how listeners would deal with the potentially conflicting contextual information.

Furthermore, participants in the current study were put under time pressure to respond quickly, and reaction time (RT) was measured in addition to the traditional accuracy measure. RT is a commonly used behavioral measure in experimental investigations of spoken language comprehension, including studies on the lexical processing of tone (Chen & Cutler, 1996; Cutler & Chen, 1995, 1997; Lee, 2007; Ye & Connine, 1999; Yip, 2001; Yip, Leung, & Chen, 1998). The motivation for using RT was to provide a corroborating and potentially more sensitive response measure that may tap into the “online” nature of speech processing. In particular, it has been argued that tone responses made under time pressure reflect true perceptual processes and are relatively free from response strategies or post-perceptual processes (Cutler, Dahan, & van Donselaar, 1997). Given the temporally brief nature of some of the stimuli, it would be of interest to examine the pattern of RT in addition to that of accuracy. Finally, compared with the relatively small number (6) of participants in Gottfried and Sutter (1997), 40 participants were used in the current study to evaluate the generalizability of the effects.

In sum, the present study attempted to address the following questions: How well can native listeners identify Mandarin tones in intact, center-only, silent-center, and onset-only syllables? Are there differences among the four tones in identification accuracy and RT? Is tone identification influenced by the match or mismatch between the carrier tone and the test tone? How well are these syllables identified when presented in isolation, in the original carrier, and in a cross-spliced context? These issues were investigated in a pretest and three experiments. The pretest evaluated baseline tone identification performance in the speeded-response task. Experiment 1 examined tone identification in syllables excised from carrier phrases. Experiment 2 looked at tone identification in syllables with original carriers. Experiment 3 investigated tone identification in syllables in the cross-spliced context.

2. Pretest

In the pretest, monosyllabic Mandarin words were read in isolation and presented to native listeners without acoustic modification. The pretest served three purposes. First, it was intended to familiarize the participants with the experimental task, since responding to auditory stimuli by pressing buttons under time pressure in a laboratory setting could be quite unnatural for some participants. Second, it was conducted to make sure the stimuli read by the speaker were indeed intelligible to the participants before they proceeded to the actual experiments. Finally, since the stimuli were read in isolation, response accuracy and RT to these citation form stimuli could serve as a baseline for comparison to the performance in subsequent experiments.

2.1. Method

2.1.1. Materials

Twelve minimal tone pairs were selected including all six tonal contrasts in Mandarin: 1–2 (xing 星行, he 合), 1–3 (xi 西洗, ma 妈马), 1–4 (si 司四, gao 高告), 2–3 (wan 完晚, you 由有), 2–4 (shi 十是, ren 人任), and 3–4 (da 打大, mei 美妹). All of these are high frequency, common words known by all the participants, who are familiar with the convention of designating Mandarin tones by numbers 1–4. Ideally, minimal tone pairs would be selected from the same set of syllables to control for segmental structure, considering the potential effects of syllabic structure and consonantal onset on $f_0$ (Shih, 1987; Xu & Xu, 2003). However, since the same set of stimuli would be presented to a group of non-native listeners to compare the native and
non-native performance directly (Lee, Tao, & Bond, 2006), our primary concern was that the selected words should have been learned and known by the non-native learners as well. While lexical status and word frequency/familiarity have been shown to impact phonetic perception (Connine, Clifton, & Cutler, 1987; Ganong, 1980; McQueen, 1991), including tone identification (e.g., Fox & Unkefer, 1985), there is no conclusive perceptual evidence of syllabic structure or consonant voicing affecting tone identification. Nonetheless, efforts were made to select words with simple syllabic structure and to include both voiced and voiceless onsets whenever possible. In the end, all selected words were CV or CV-nasal syllables. The distribution of onset voicing (voiced/voiceless) was 1 voiced/5 voiceless for Tone 1, 3/3 for Tone 2, 4/2 for Tone 3, and 2/4 for Tone 4.

The 24 syllables were read in isolation by the second author, a phonetically trained female native speaker of Mandarin. The recording was made in a sound-treated booth in the School of Hearing, Speech and Language Sciences at Ohio University with a high-quality microphone (Audio-technica AT825 field recording microphone) connected through a preamplifier and A/D converter (USBPre microphone interface) to a Windows personal computer (Dell). The recording was sampled using the Brown Lab Interactive Speech System (BLISS, Mertus, 2000) at 20 kHz with 14-bit quantization. Each syllable was identified from the BLISS waveform display, excised from the master file, and saved as an individual audio file.

2.1.2. Participants

Forty adult native speakers of Mandarin (age: 22–35) participated in the pretest and subsequent experiments. They were recruited from the Ohio University community and were paid $10 for their participation. All spoke Mandarin on a daily basis and none reported any speech or hearing difficulties. Participants came from various geographical regions in China, but all identified Mandarin as their native language.

2.1.3. Procedure

The 24 syllables, saved as individual audio files, were imported to AVRunner, the subject-testing program in BLISS, for stimulus presentation. For each participant, AVRunner assigned a uniquely randomized presentation order such that no two participants received the same order of presentation. Participants were tested individually or in pairs in a quiet room in the Department of Linguistics at Ohio University. They listened to the syllables through a pair of high-quality headphones (Koss R80) and were instructed to identify the tone of each syllable by pressing buttons labeled “1”, “2”, “3”, and “4” on a computer keyboard, representing the four Mandarin tones, respectively. They were also instructed to make a response as soon as possible without sacrificing accuracy.

2.1.4. Data analysis

Acoustic analyses were conducted with the Mev and the Pitch programs in BLISS. Mev displays the waveform. Pitch estimates fundamental frequency ($f_0$) via autocorrelation and displays the $f_0$ contour of an utterance. Response accuracy and RT were automatically recorded by BLISS. RT was measured from stimulus offset to avoid the potential confound of intrinsic duration differences among the tones. One-way repeated measures analyses of variance (ANOVAs) were conducted on accuracy and RT with stimulus tone (1, 2, 3, and 4) as the fixed factor and participants as the random factor. Only correct responses were included in the RT analysis. When a main effect from the ANOVAs was significant, the Bonferroni post-hoc test was used for pair-wise means comparisons to keep the family-wise Type I error rate at 5%.

2.2. Results

2.2.1. Acoustic analysis

The duration of the syllables was measured from the waveform display. Table 1 shows the duration of the syllables divided into onset and rhyme. For syllables with a voiceless onset, the rhyme was defined as the onset of acoustic periodicity. For syllables with a voice onset, the rhyme was defined as the start of the vowel, which was identified by consulting both the waveform and the spectrogram. The average duration of the rhyme was 317 ms (SD = 35) for Tone 1, 441 ms (SD = 33) for Tone 2, 488 ms (SD = 46) for Tone 3, and 329 ms
(SD = 27) for Tone 4. A one-way ANOVA showed a significant effect of tone: $F(3, 20) = 32.64, p < .0001$. The Bonferroni post-hoc test showed all pair-wise comparisons were significant except between Tones 1 and 4 and between Tones 2 and 3. In other words, Tones 2 and 3 were longer than Tones 1 and 4. This pattern is consistent with previous acoustic measurements of isolated Mandarin tones (e.g., Tseng, 1981).

Fig. 1 shows the $f_0$ contours of the 24 syllables used in the pretest, arranged by the four Mandarin tones. The four panels from left to right show Tone 1, Tone 2, Tone 3, and Tone 4, respectively. The start of $f_0$ data points was aligned to time 0 in the figure.

### 2.2.2. Accuracy

Tone identification accuracy for these isolated syllables is shown in Table 2. On average, the accuracy exceeded 90% for all four tones. While Tone 4 appears to be identified most accurately and Tone 3 least accurately, the ANOVA did not show any statistically significant differences among the four tones: $F(3, 117) = 2.58, p > .05$.

The overall high accuracy validated the intelligibility and usefulness of the speaker’s speech. Since the stimuli were read in isolation and presented to the native listeners in the quiet, one would expect the response

### Table 1

<table>
<thead>
<tr>
<th>Tone pair</th>
<th>Syllable</th>
<th>Onset duration</th>
<th>Rhyme duration</th>
<th>Syllable</th>
<th>Onset duration</th>
<th>Rhyme duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–2</td>
<td>xing1</td>
<td>197</td>
<td>336</td>
<td>xing2</td>
<td>189</td>
<td>451</td>
</tr>
<tr>
<td></td>
<td>he1</td>
<td>154</td>
<td>350</td>
<td>he2</td>
<td>109</td>
<td>423</td>
</tr>
<tr>
<td>1–3</td>
<td>xil</td>
<td>198</td>
<td>323</td>
<td>Xi3</td>
<td>233</td>
<td>457</td>
</tr>
<tr>
<td></td>
<td>ma1</td>
<td>98</td>
<td>305</td>
<td>ma3</td>
<td>150</td>
<td>457</td>
</tr>
<tr>
<td>1–4</td>
<td>si1</td>
<td>223</td>
<td>252</td>
<td>Si4</td>
<td>262</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>gao1</td>
<td>17</td>
<td>336</td>
<td>gao4</td>
<td>13</td>
<td>344</td>
</tr>
<tr>
<td>2–3</td>
<td>wan2</td>
<td>70</td>
<td>476</td>
<td>wan3</td>
<td>98</td>
<td>561</td>
</tr>
<tr>
<td></td>
<td>you2</td>
<td>91</td>
<td>438</td>
<td>you3</td>
<td>102</td>
<td>513</td>
</tr>
<tr>
<td>2–4</td>
<td>shi2</td>
<td>244</td>
<td>387</td>
<td>shi4</td>
<td>238</td>
<td>309</td>
</tr>
<tr>
<td></td>
<td>ren2</td>
<td>53</td>
<td>473</td>
<td>ren4</td>
<td>145</td>
<td>372</td>
</tr>
<tr>
<td>3–4</td>
<td>da3</td>
<td>9</td>
<td>497</td>
<td>da4</td>
<td>7</td>
<td>330</td>
</tr>
<tr>
<td></td>
<td>mei3</td>
<td>103</td>
<td>440</td>
<td>mei4</td>
<td>119</td>
<td>320</td>
</tr>
</tbody>
</table>
accuracy to be perfect. A number of reasons could have contributed to the less-than-perfect identification. First is the temporally demanding nature of the speeded-response task. In particular, participants were under time pressure to make a response as quickly as possible, which could have compromised their response accuracy. In addition, despite the listeners’ native knowledge of tones, making judgments by the tone names (1, 2, 3, and 4) may not be completely natural in the experimental setting. Indeed, further inspection of individual participant data revealed that one participant had only one correct response out of the 24 possible responses. When this participant was excluded from the data analysis, the average response accuracy improved to 96%, 95%, 93%, and 100% for the four tones, respectively. Further instructions were provided to this participant to make sure the experimental procedure was clearly understood and followed in subsequent experiments.

2.2.3. Reaction time

Table 2 also shows the average RT of tone identification. As noted earlier, RT was measured from stimulus offset to avoid the potential confound of intrinsic duration differences among the four tones. Note that the RT adjustment was made for individual stimulus items, not based on the average durations reported in the acoustic analysis results. On average, Tone 4 took the least and Tone 1 took the most amount of time to identify. A one-way repeated measures ANOVA revealed a significant main effect of tone: \( F(3, 117) = 3.06, p < .05 \). Pair-wise means comparisons by the Bonferroni test showed significant differences between Tones 1 and 3 and between Tones 1 and 4. That is, Tone 1 needed more time to identify compared with Tones 3 and 4. Since RT was measured from the syllable offset and the longest syllables were not responded to most quickly, the observed difference is unlikely due to the intrinsic duration difference among the four tones. The only potential reason for the observed difference is that Tone 1 is the only non-contour tone among the four tones when produced in isolation. Perhaps, it was generally easier to monitor tones when there are \( f_0 \) movements. When such a movement is lacking in a non-contour tone, it took more time to make the tone judgment. Whether this contributes to the slower reaction awaits further research.

2.2.4. Summary

The high accuracy of tone identification in the pretest validated the speech recorded by the speaker for this study. The results showed that all four tones were identified with comparable accuracy. The RT analysis, however, showed that Tones 3 and 4 were identified more quickly than Tone 1. While the reason for this difference remains speculative, this result showed that the latency measure may reveal potential processing differences that may not be detected by the traditional accuracy measure.

In the next three experiments, acoustic modifications were introduced to the syllables, which were presented in various phonetic contexts. Specifically, Experiment 1 examined tone identification in syllables excised from the original carrier phrase and presented in isolation. In Experiment 2, the syllables were presented with the original carrier phrases. In Experiment 3, the syllables were excised from one carrier phrase and cross-spliced with the other carrier phrase. The effects of acoustic modification and phonetic context on the speed and accuracy of tone identification were evaluated.
3. Experiment 1

In Experiment 1, the syllables selected as stimuli in the pretest were recorded again, this time in two carrier phrases. These test syllables were digitally excised from the carrier phrases, modified to generate intact, center-only, silent-center, and onset-only syllables, and presented to listeners for tone identification. As noted in the introduction, the research questions included: (1) How well do listeners recognize lexical tones when only partial acoustic information is available? (2) How well do listeners recognize lexical tones without the original carrier phrase? (3) Does the match or mismatch between the offset of the carrier tone and the onset of the test tone have an effect on tone identification?

3.1. Method

3.1.1. Materials

The 24 words used in the pretest were selected for stimulus construction in this and subsequent experiments. A new recording was made by the same speaker with the same recording procedure. In contrast to the pretest, where the syllables were read in isolation, the stimuli were now read with two carrier phrases to examine the effects of phonetic context. In particular, the syllables were recorded with the carrier phrases Qing3 shuo1 __ (“Please say __”) and Qing3 kan4 __ (“Please look at __”) such that the offset of the carrier phrase and the onset of the test syllable were either a match or mismatch in \( f_0 \). For example, the last syllable of the carrier phrase shuo1 and the test syllable xing1 were a “match” in \( f_0 \) because shuo1 ends in a high \( f_0 \) and xing1 also starts with a high \( f_0 \). In contrast, the carrier shuo1 and the test xing2 were a “mismatch” in \( f_0 \) because shuo1 ends in a high \( f_0 \) and xing2 starts with a low \( f_0 \). Since one of the two carrier phrases ends in a high \( f_0 \) and the other ends in a low \( f_0 \), and since two of the four test tones begin with a high \( f_0 \) and the other two begin with a low \( f_0 \), the number of “match” cases is equal to that of “mismatch” cases.

It should be noted that the test syllables were excised from the carrier phrases and presented in isolation even though they were originally recorded with the carrier phrases. To that end, each test syllable was identified from the waveform display, excised from the carrier phrases, and saved as individual audio files for stimulus presentation. Each test syllable was then digitally modified with BLISS to generate four types of syllables: intact, center-only, silent-center, and onset-only. Following Gottfried and Suiter (1997), the first six and final eight pitch periods of the intact syllables were digitally edited. In particular, center-only syllables were constructed by removing the first six and final eight pitch periods of a syllable; silent-center syllables were generated by preserving only the first six and final eight pitch periods; and onset-only syllables were produced by preserving only the first six pitch periods. The removed part or parts were digitally “silenced” such that the overall duration remained the same as that of the intact syllables. There were no perceptible clicks as a result of the signal processing; therefore, no further tapering procedure was applied. An example of the acoustic modifications is shown in Fig. 2. A total of 192 stimuli (4 tones × 6 tokens per tone × 4 modifications × 2 carrier phrases) were used in this experiment.

3.1.2. Participants

The 40 participants in the pretest also participated in this and subsequent experiments.

3.1.3. Procedure

The experimental procedure was identical to that used in the pretest, with the exception that 192 stimuli were used in the present experiment. The experiment was administered following a 5-min break after the participants completed the pretest.

3.1.4. Data analysis

Acoustic analyses were conducted of the duration and \( f_0 \) of the syllables. In addition, response accuracy and RT were recorded by BLISS automatically in this and subsequent experiments. As noted earlier, RT was measured from stimulus offset to avoid the potential confound of intrinsic duration differences among the tones. It should also be noted that the RT adjustment was made individually for each stimulus item based on the new recording (which were read with the carrier phrases), not the recording used in the pretest.
Repeated measures ANOVAs were conducted on response accuracy and RT with acoustic modification (intact, center-only, silent-center, onset-only), stimulus tone (1, 2, 3, and 4), and preceding tone (i.e., whether the offset of the carrier and the onset of the test tone were a match or mismatch in $f_0$) as fixed factors and participants as the random factor. Only correct responses were included in the RT analysis. When a main effect was significant, the Bonferroni post-hoc test was used for pair-wise means comparisons to keep the family-wise Type I error rate at 5%.

### 3.2. Results

#### 3.2.1. Acoustic analysis

For each syllable used in this and subsequent experiments, duration of four components was measured including (a) the onset consonant, (b) the first six pitch periods of the rhyme, (c) the center of the rhyme, and (d) the final eight pitch periods of the rhyme. For syllables with a voiceless onset, the rhyme was defined as the onset of acoustic periodicity. For syllables with a voice onset, the rhyme was defined as the start of the vowel, which was identified by consulting both the waveform and the spectrogram. In terms of acoustic modification, parts a, b, and d were silenced for center-only syllables; part c was silenced for silent-center syllables; parts c and d were silenced for onset-only syllables.

To evaluate the duration of the components relevant to the tone identification task, two-way ANOVAs were conducted with tone and preceding tone as fixed factors on the duration of the first six periods (part b), the center (part c), the final eight periods (part d), and the total duration of the syllable. Table 3 shows the average duration of these components as a function of tone and preceding tone. For the duration of the first six

---

**Table 3**

Average duration (ms) of various components of the syllables used in Experiments 1–3

<table>
<thead>
<tr>
<th>Preceding tone</th>
<th>Tone</th>
<th>First six periods</th>
<th>Center</th>
<th>Final eight periods</th>
<th>Total duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>19 (1)</td>
<td>217 (23)</td>
<td>26 (2)</td>
<td>393 (52)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>31 (1)</td>
<td>280 (60)</td>
<td>26 (1)</td>
<td>464 (15)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>36 (5)</td>
<td>344 (55)</td>
<td>37 (1)</td>
<td>516 (23)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17 (1)</td>
<td>200 (35)</td>
<td>64 (12)</td>
<td>379 (77)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>18 (1)</td>
<td>204 (21)</td>
<td>28 (4)</td>
<td>384 (58)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>32 (2)</td>
<td>271 (50)</td>
<td>28 (4)</td>
<td>446 (21)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>39 (8)</td>
<td>323 (49)</td>
<td>42 (4)</td>
<td>518 (64)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>18 (1)</td>
<td>204 (28)</td>
<td>53 (9)</td>
<td>385 (72)</td>
</tr>
</tbody>
</table>

Standard deviation is shown in parenthesis.

---

Fig. 2. Acoustic modifications conducted in Experiment 1, illustrated by the syllable *gao4*. From top down: intact, center-only, silent-center, and onset-only syllables. See text for details.
periods, the main effect of tone was significant, $F(3, 40) = 93.23, p < .0001$. Post-hoc tests showed that all pair-wise comparisons were significant except for the difference between Tones 1 and 4. On average, Tones 1 and 4 had shorter onset durations than Tone 2, whose onset duration was in turn shorter than Tone 3. This implies that Tones 1 and 4 had higher initial $f_0$'s than Tones 2 and 3, which is consistent with the $f_0$ pattern of tones produced in isolation.

For the duration of the center, the main effect of tone was significant, $F(3, 40) = 25.2, p < .0001$. Post-hoc tests showed that all pair-wise comparisons were significant except for the difference between Tones 1 and 4. On average, Tones 1 and 4 had shorter center durations than Tone 2, whose center duration was in turn shorter than Tone 3. For the duration of the final eight periods, the main effect of tone was significant, $F(3, 40) = 74.07, p < .0001$. Post-hoc tests showed that all pair-wise comparisons were significant except for the difference between Tones 1 and 2. On average, Tones 1 and 2 had shorter final durations than Tone 3, whose final duration was in turn shorter than Tone 4. This implies that the Tones 1 and 2 had higher final $f_0$'s than Tones 3 and 4, which is consistent with the $f_0$ pattern of tones produced in isolation. In addition, the tone–preceding tone interaction was also significant, $F(3, 40) = 3.97, p < .05$. The interaction arose from the longer final duration for Tone 4 when following the Tone 1 carrier phrase. Finally, for total duration, the main effect of tone was significant, $F(3, 40) = 17.11, p < .0001$. Post-hoc tests showed that all pair-wise comparisons were significant except for the difference between Tones 1 and 4. The difference in intrinsic duration highlighted the need to control for this factor in the RT analyses.

Fig. 3 shows the $f_0$ contours of the syllables used in this and subsequent experiments, arranged by tone and preceding tone. The start of $f_0$ data points was aligned to time 0 in the figure. These $f_0$ contours, generated by an autocorrelation algorithm, appear to closely resemble those produced in isolation. The only visible differences between these excised tones and the isolated tones reported in the pretest (Fig. 1) are that the excised Tone 3 shows more variability in the onset $f_0$, and that the excised Tone 4 has higher onset $f_0$ than the isolated Tone 4. Other than these contrasts, the overall $f_0$ shapes between these two figures seem to be largely similar to each other.

To obtain a more fine-grained measure of the $f_0$ information available in the partial acoustic input, the first six pitch periods for each syllable were identified from the waveform display and the $f_0$'s during this interval were manually measured for each glottal cycle. To that end, each glottal cycle was manually marked on the

![Fig. 3. The $f_0$ contours of the 24 syllables used in Experiment 1, arranged by test tone and preceding tone in the carrier phrases. The four panels from left to right show Tone 1, Tone 2, Tone 3, and Tone 4, respectively. The top panels show test tones produced with the carrier phrase Qing3shuo1; the bottom panels show test tones produced with the carrier phrase Qing3kan4. The start of $f_0$ data points was aligned to time 0 in the figure.](image)
waveform display and the $f_0$ values were recorded cycle by cycle. Fig. 4 shows these first six $f_0$ values of the four tones with each data point averaged across six stimulus tokens. This figure shows that the four tones form two distinct groups, with Tones 1 and 4 showing higher $f_0$'s and Tones 2 and 3 showing lower $f_0$'s. In addition, all four contours are fairly flat, showing little if any contrast in $f_0$ shape.

To obtain a quantitative measure of the tone differences for each tone, a mean $f_0$ value was calculated by averaging across the six data points. For syllables with the Tone 1 carrier, the average $f_0$ was highest for Tones 4 ($359 \text{ Hz, SD = 22}$) and 1 ($335 \text{ Hz, SD = 9}$), followed by Tone 2 ($194 \text{ Hz, SD = 9}$), and lowest for Tone 3 ($173 \text{ Hz, SD = 22}$). For syllables with the Tone 4 carrier, the average $f_0$ was also highest for Tones 4 ($340 \text{ Hz, SD = 14}$) and 1 ($328 \text{ Hz, SD = 7}$), followed by Tone 2 ($187 \text{ Hz, SD = 12}$), and lowest for Tone 3 ($161 \text{ Hz, SD = 27}$). This pattern corroborates with the duration analysis reported earlier. A two-way ANOVA was conducted of the average $f_0$ with tone and preceding tone as fixed factors. The ANOVA showed a main effect of tone, $F (3, 40) = 375.71$, $p < .0001$. Post-hoc tests showed that all pair-wise comparisons were significant except for the difference between Tones 1 and 4. In addition, there was a main effect of preceding...
Specifically, the average $f_0$ of the first six pitch periods was 11 Hz higher when the syllables followed the Tone 1 carrier compared with the Tone 4 carrier. There was no tone–preceding tone interaction, indicating that the difference was uniform across the four tones. This moderate yet statistically significant difference is what would be expected from carryover coarticulation (Xu, 1997). In particular, since Tone 1 ends at a higher $f_0$ than Tone 4, its higher ending $f_0$ “spilled over” to influence the starting $f_0$ of the test tone.

3.2.2. Accuracy

Fig. 5 shows tone identification accuracy by acoustic modification and tone. On average, identification accuracy was highest for intact (97%) and center-only (94%) syllables, followed by silent-center syllables (86%), and lowest for onset-only syllables (73%). A repeated measures ANOVA revealed a main effect of modification: $F (3, 117) = 182.92, p < .0001$. All pair-wise comparisons were significant ($p < .0001$) except for the difference between intact and center-only syllables. This result contrasts with Gottfried and Suiter (1997, Experiment 2), who found no difference among intact, center-only, and silent-center syllables. That is, the silent-center syllables in the current experiment were not identified as accurately as the intact and center-only syllables. Importantly, all the accuracy values clearly exceeded chance-level performance (25%). Even the onset-only syllables were identified quite accurately when presented in isolation (73%, as opposed to 55% in Gottfried & Suiter, 1997). Specifically, how this might have been achieved will be discussed shortly.

There was also a main effect of tone: $F (3, 117) = 44.13, p < .0001$. On average, identification accuracy was highest for Tone 4 (97%), followed by Tones 3 (90%) and 1 (87%), and lowest for Tone 2 (76%). All pair-wise comparisons of accuracy were significant ($p < .005$) except for the difference between Tones 1 and 3. The tone effect, however, was not uniform across modifications. There was a significant interaction between tone and modification: $F (9, 351) = 38.61, p < .0001$. It can be seen from Fig. 5 that Tones 1 and 2 were particularly compromised in the onset-only syllables, so was Tone 2 in the silent-center syllables. In contrast, Tones 3 and 4 were relatively spared despite the modifications. From these results, it appears that Tone 2 identification may rely more on syllable center information and that Tone 1 identification could benefit from the presence of both onset and offset.

No main effect of preceding tone was found. That is, for a given syllable, identification accuracy did not differ whether it was embedded in one carrier phrase or the other. That is, the acoustic difference found in the $f_0$ analysis did not appear to make an impact on perception. Nonetheless, there was a significant interaction among tone, modification, and preceding tone: $F (9, 351) = 3.18, p < .005$. For Tone 4 in onset-only syllables,
a matching preceding tone produced higher accuracy. In contrast, for Tone 3 in center-only syllables, a
mismatching preceding tone generated higher accuracy. Presumably, a mismatching preceding tone would
have introduced more deviation from the canonical contour to the test tone. It is not clear why a mismatching
preceding tone would facilitate identification, but the effect was limited to Tone 3 stimuli and did not
generalize to other tones.

3.2.3. Confusion patterns
To explore the types of tone errors made in the identification task, responses were tabulated by tone. Since
the accuracy for the intact, center-only, and silent-center syllables were fairly high (all over 86%), the
confusion matrix analysis was conducted only for the onset-only condition. Specifically, a confusion matrix
was generated for each participant; the numbers in the cells were then converted to percentages before being
averaged across participants. Table 4 shows the error pattern. Tone 1 was primarily confused with Tone 4, and
Tone 2 was predominantly confused with Tone 3. Broadly speaking, this pattern is consistent with the acoustic
analyses reported earlier that Tones 1 and 4 shared a similar \( f_0 \) onset in the lower 300 Hz range and that Tones
2 and 3 shared a similar \( f_0 \) onset in the upper 100 Hz range. The confusion pattern is also consistent with what
has been reported in previous tone identification studies (Blicher, Diehl, & Cohen, 1990; Gottfried & Suiter,

This perceptual error pattern, however, could not be completely predicted from the acoustic data. In
particular, recall that the mean \( f_0 \) averaged over the first six pitch periods was highest for Tones 1 and 4
(no statistical difference), followed by Tone 2, and lowest for Tone 3. In addition, the \( f_0 \) contours during these
six pitch periods were flat for all four tones and showed no distinct patterns. Since the only available acoustic
information for the onset-only syllables was from the first six pitch periods, the acoustic data would predict
more confusion between Tones 1 and 4 than between Tones 2 and 3. However, the actual error pattern
appears be the opposite. In addition, the acoustic data could not predict the asymmetries seen in the confusion
pattern. In particular, Tone 1 was more often misidentified as Tone 4 than vice versa; Tone 2 was more often
misidentified as Tone 3 than vice versa. It is not clear how the acoustic data could account for this asymmetry.
It should be noted that data from Gottfried and Suiter (1997, Experiment 2) also showed dominant Tone 3
responses, although their listeners preferred Tones 1–4 instead.

How do we account for the asymmetries in the confusion pattern? It is possible that upon hearing a low \( f_0 \)
onset, listeners would give a Tone 3 response unless they heard positive evidence of a rising contour for Tone
2. This interpretation is also consistent with the result that Tone 2 identification was particularly compromised
in the silent-center and onset-only condition. Specifically, previous acoustic studies and perceptual tests
(Shen & Lin, 1991; Shen, Lin, & Yan, 1993) showed that Tones 2 and 3, when produced in citation form, differ
in the \( f_0 \) “turning point”, or where the \( f_0 \) contour turns from falling to rising. This turning point normally

Table 4
Confusion matrix of tone identification responses in onset-only syllables in Experiment 1

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone 1</td>
<td>68</td>
<td>3</td>
<td>2</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Tone 2</td>
<td>4</td>
<td>46</td>
<td>47</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tone 3</td>
<td>3</td>
<td>7</td>
<td>87</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Tone 4</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td><strong>G&amp;S (1997)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone 1</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Tone 2</td>
<td>12</td>
<td>28</td>
<td>59</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tone 3</td>
<td>16</td>
<td>17</td>
<td>65</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Tone 4</td>
<td>66</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

Numbers shown are percentages. Top: data from current experiment. Bottom: data from Gottfried and Suiter (1997), reproduced here for comparison.
shows up in the syllable center and would be missing from the silent-center and onset-only syllables in the current study, which would explain why Tone 2 was identified particularly poorly from the these two types of syllables. As for the bias towards Tone 4 in the Tones 1–4 confusion, it is possible that upon hearing a high \( f_0 \) onset, listeners had a tendency to give a Tone 4 response unless they had evidence of a flat contour for Tone 1. This interpretation is consistent with the result that Tone 1 identification was particularly compromised when the center and offset was missing. It should be noted that Gottfried and Suiter (1997) found the opposite pattern for the Tones 1–4 confusion, i.e., Tone 1 response dominated the confusion. The interpretation was that without the following \( zi \) syllable, listeners could no longer use the low \( f_0 \) onset that was present in the \( zi \) syllable to infer the low \( f_0 \) ending of Tone 4; thus, when the listeners heard a high onset for the test syllable, they would give a Tone 1 response. These conflicting findings will need to be addressed by further research.

How do we account for the better-than-chance accuracy of tone identification from merely six pitch periods in syllables presented in isolation? Since the average amount of acoustic input ranged from only 17 to 39 ms (Table 3) and the first six periods for all four tones showed virtually flat \( f_0 \) contours (Fig. 4), it is unlikely that the listeners utilized \( f_0 \) shape to identify the tones. From the confusion data, it appears as if the listeners had a tendency to give a Tone 4 response when they heard a high \( f_0 \) onset, and to give a Tone 3 response when they heard a low \( f_0 \) onset. Taken together, it looks like \( f_0 \) height might have played a more prominent role than \( f_0 \) contour for tone identification from limited acoustic input. However, the use of \( f_0 \) height implies the listener’s knowledge of the speaker’s \( f_0 \) range. That is, to judge whether a particular \( f_0 \) value is high or low, a listener has to be able to know the speaker’s \( f_0 \) range as a reference frame. Since \( f_0 \) range varies across speakers, it makes intuitive sense that listeners have to engage in some kind of speaker normalization when perceiving tones, just as they do for vowels (e.g., Ladefoged & Broadbent, 1957).

Indeed, several studies have shown that the same \( f_0 \) contour could be identified as different tones depending on the perceived \( f_0 \) range of the speaker (Leather, 1983; Lin & Wang, 1984; Moore & Jongman, 1997). However, it is not known from these studies how much acoustic input is needed for the speaker normalization to be achieved. The result from the current experiment that tones could be identified reliably with six pitch periods presented in isolation seems to suggest that not much is needed, even in the absence of an immediate phonetic context. Since only one speaker was used in the present study, the listeners had plenty of opportunities to estimate the speaker’s \( f_0 \) range from exposure to the speaker’s speech in the pretest and the current experiment, which could have contributed to their successful judgment of \( f_0 \) height within the speaker’s \( f_0 \) range, and thus the high accuracy of tone identification.

### 3.2.4. Reaction time

Fig. 5 shows the RT of tone identification by acoustic modification and tone. On average, identification of the intact syllables took the least amount of time (644 ms), followed by the silent-center (770 ms), center-only (797 ms), and onset-only (821 ms) syllables. A repeated measures ANOVA revealed a main effect of tone: \( F(3, 117) = 30.71, p < .0001 \). The three pair-wise comparisons involving intact syllables were significant \( (p < .0001) \). That is, the RT for intact syllables was shortest and no difference was found among the remaining three types of modified syllables. These results corroborated with the accuracy analysis. In particular, a faster response to the intact syllables, like the higher accuracy reported earlier, was to be expected. The little increase in RT among the three modified conditions indicated that the accuracy contrasts shown earlier did not arise from any speed-accuracy tradeoff. Finally, although the center-only syllables were identified as accurately as the intact syllables, the RT analysis revealed that they were not treated in the same way as intact syllables in online speech processing. The accuracy measure alone would not have revealed this pattern.

There was also a main effect of tone: \( F(3, 117) = 24.169, p < .0001 \). The RTs for Tones 4 (634 ms) and 3 (639 ms) were shorter than those for Tones 1 (872 ms) and 2 (886 ms). Pair-wise means comparisons showed significant differences between Tone 1 and Tones 3 and 4, as well as between Tone 2 and Tones 3 and 4. That is, RT was shorter for Tones 3 and 4 compared with Tones 1 and 2. Again, this pattern is consistent with the accuracy analysis showing Tones 3 and 4 were relatively spared by the modifications. Nonetheless, the effect of modification was not equal on all four tones, as indicated by a significant interaction between tone and modification: \( F(9, 351) = 38.61, p < .0001 \). As Fig. 5 shows, while the intact syllables always took the least amount of time to identify for all four tones, the effect of modifications showed some variation by tone.
Compared with the center-only syllables, silent-center syllables further compromised Tone 2 only, and onset-only syllables compromised Tones 1 and 2 while having relatively little impact on Tones 3 and 4. This pattern is again consistent with the accuracy analysis showing that Tone 2 identification relies more on syllable center and that Tone 1 could benefit from the presence of both onset and offset.

Finally, as was in the accuracy analysis, no main effect of preceding tone was found for RT either. The speed of response did not differ whether the tone was embedded in one carrier phrase or another. There were no other effects.

3.2.5. Summary

In this experiment, acoustic modifications were introduced to Mandarin syllables that were recorded with carrier phrases but were excised and presented in isolation. Acoustic analyses on the duration of these excised syllables showed that the four tones differed in onset, center, offset, and total duration, as would be expected from tones produced in isolation. Acoustic analyses on $f_0$ also showed that these excised syllables were similar to the tones produced in isolation with distinct $f_0$ shapes and onset $f_0$’s. There was a small but significant effect of preceding tone on the onset $f_0$ of the test tone, indicating a carryover effect from the preceding tone. However, the acoustic difference did not impact identification accuracy or RT.

Not surprisingly, both the accuracy and the RT data indicated that the intact syllables were the easiest to identify. The center-only syllables were identified as accurately as the intact syllables, but the use of the RT measure revealed that they were not processed as fast. For the modified syllables, identification accuracy declined as the amount of acoustic input decreased. However, the removal of acoustic information in the modified syllables did not compromise identification accuracy anywhere to the chance level. The little decrement in RT among the three modifications further indicated that the accuracy differences were free from speed-accuracy tradeoff and reflected true online processing effects. Although Tone 2 identification was the least accurate and took the longest time, the tone–modification interaction revealed that not all tones were treated equally by the modifications. The accuracy and RT data jointly suggest that Tone 2 identification was particularly hampered by the absence of the syllable center and that Tone 1 could benefit from the presence of both syllable onset and offset. In contrast, the identification of Tones 3 and 4 did not appear to be significantly compromised by the modifications.

The confusion pattern analysis on the onset-only syllables showed the Tone 2–Tone 3 confusion commonly reported in the literature, and substantial Tone 1–Tone 4 confusions. Interestingly, the confusions were not symmetric. Our listeners preferred Tone 3 to Tone 2 and Tone 4 to Tone 1. It was speculated that with such limited acoustic input from the syllables presented in isolation, the intrinsic $f_0$ shape information would not be available or useful. Instead, listeners might have managed to estimate the speaker’s $f_0$ range, based on which they were able to evaluate the $f_0$ height of the acoustic input for accurate tone identification.

4. Experiment 2

Experiment 1 showed that acoustically modified tones were identified with high accuracy by native listeners even when the majority of a syllable was silenced. Remarkably, the syllables were all presented in isolation, meaning that the listeners would have to rely on syllable-intrinsic information to judge tone identity. In this experiment, tone identification from limited acoustic input was evaluated when the tones were presented in context. Recall that the syllables used in Experiment 1 were originally recorded with two carrier phrases. The offset $f_0$ of the carrier tone was either a match or a mismatch to the onset $f_0$ of the test syllable. In Experiment 2, the test syllables would be presented with the original carrier phrases.

What will be the effect of the presence of a context on tone identification? Intuition and the literature both suggest that a context would aid speech perception because it provides additional syllable-extrinsic information that may be relevant to tone identification. As noted in the Introduction, acoustic and perceptual studies of tone have implicated the role of tonal context, although there are conflicting findings regarding the direction and magnitude of tonal coarticulation. The acoustic analyses reported in Experiment 1 showed some evidence of carryover coarticulation, i.e., the onset $f_0$ of the test tone varied as a function of the offset $f_0$ of the preceding tone. Although the acoustic contrast did not affect tone identification when the tones were excised
and presented in isolation, the pattern of tone identification response could change when the context is actually present.

4.1. Method

The stimuli, participants, procedure, and data analyses in this experiment were identical to those used in Experiment 1 except that the test syllables were now presented with the original carrier phrases in this experiment. This experiment was administered following a 5-min break after the participants completed Experiment 1.

4.2. Results

4.2.1. Accuracy

Fig. 6 shows tone identification accuracy by acoustic modification and tone. On average, identification was most accurate for intact (95%) and center-only syllables (95%), followed by silent-center syllables (89%), and least accurate for onset-only syllables (80%). A repeated measures ANOVA revealed a main effect of modification: $F(3, 117) = 99.37$, $p < .0001$. All pair-wise comparisons of acoustic modification were significant ($p < .0001$) except for the difference between intact and center-only syllables. This pattern is identical to what was found in Experiment 1 but contrasts with Gottfried and Suiter (1997, Experiment 1), who found no difference among the intact, center-only, and silent-center syllables.

In addition, there was a main effect of tone: $F(3, 117) = 21.54$, $p < .0001$. On average, identification accuracy was higher for Tone 4 (96%), Tone 3 (92%), and Tone 1 (91%) than Tone 2 (80%). The order of accuracy for the tones was identical to what was found in Experiment 1, although in this experiment only the three pair-wise means comparisons involving Tone 2 were significant ($p < .0001$). The tone effect was not uniform across modifications, as indicated by a significant interaction between tone and modification: $F(9, 351) = 16.29$, $p < .0001$. As in Experiment 1, Tone 2 was particularly compromised in the silent-center syllables, and Tones 1 and 2 were identified more poorly, compared with Tones 3 and 4, in the onset-only syllables. Finally, no main effect was found of the preceding tone in the carrier phrases. In sum, the accuracy data in this experiment showed the same pattern as what was found in Experiment 1.
4.2.2. Confusion patterns

Confusion matrices were generated only for the onset-only syllables because tone identification for the other three types of modified syllables was fairly accurate. Specifically, a confusion matrix was generated for each participant; the numbers in the cells were then converted to percentages before being averaged across participants. Table 5 shows the error patterns. As was found in Experiment 1, Tone 1 was primarily confused with Tone 4, and Tone 2 with Tone 3. Compared with Experiment 1, tone identification showed only moderate improvement for Tones 3 (2%) and 4 (1%), but the improvement was more substantial for Tones 1 (13%) and 2 (15%) due to reduced confusion with Tones 4 and 3, respectively. Consequently, the bias towards Tone 3 responses in the Tone 2–3 confusion and the bias towards Tone 4 responses in the Tone 1–4 confusion were both reduced. It appears that the presence of the context helped reducing the confusions. Table 6 also shows data from Gottfried and Suiter (1997, Experiment 1). Note that the carrier in their study was after the test syllable, while the carrier in the current study was before the test syllable. The response patterns are fairly similar in the two studies with one notable exception: Tone 2 stimuli in Gottfried and Suiter (1997) elicited primarily Tone 3 responses. In the present experiment, listeners did misidentify Tone 2 as Tone 3, but the Tone 2 response was still the majority.

4.2.3. Reaction time

Fig. 6 shows the RT for tone identification by acoustic modification and tone. On average, identification of intact syllables took the least amount of time (564 ms), followed by center-only (665 ms), silent-center (670 ms), and onset-only (694 ms) syllables. A repeated measures ANOVA revealed a significant main effect of modification: $F(3, 117) = 30.71, p < .0001$. The three pair-wise means comparisons involving intact syllables were significant ($p < .0001$). That is, the RT for intact syllables was shortest and no difference was found
among the remaining three types of syllables. This pattern is identical to what was found in Experiment 1. That is, the non-significant difference in RT among the three modified conditions indicated that the accuracy contrasts shown earlier did not arise from speed-accuracy tradeoff. Further, although the center-only syllables were identified as accurately as the intact syllables, the RT analysis showed that they were not treated in the same way as intact syllables in online tone processing.

In addition, there was a main effect of tone: \( F(3, 117) = 24.169, p < .0001 \). The RT was shorter for Tones 4 (528 ms) and 3 (560 ms) compared with Tones 1 (752 ms) and 2 (753 ms). Pair-wise means comparisons showed significant differences between Tone 1 and Tones 3 and 4, as well as between Tone 2 and Tones 3 and 4. That is, the RT is shorter for Tones 3 and 4 compared with Tones 1 and 2. These results are consistent with the accuracy analysis showing that Tones 3 and 4 were relatively spared by the modifications. This pattern is also very similar to what was found in Experiment 1 with one exception: while Tone 1 also achieved high accuracy (91%) here, the RT measure showed that it still was not responded to as fast as Tones 3 or 4. Finally, no main effect was found for the match or mismatch with the preceding tone in the carrier phrases, nor were there any interactions.

4.2.4. Summary

The findings from this experiment, where the test syllables were presented in their original carriers, showed almost identical patterns to the results from Experiment 1, where the test syllables were excised from the carriers and presented in isolation. The similarity suggests that contextual information extrinsic to the test tone did not have a significant impact, which implies that listeners relied primarily on syllable-intrinsic information for tone identification. The lack of context effect on tone identification could be attributed to the limited tonal coarticulation in the stimuli in this study. In previous acoustic investigations of tonal coarticulation, test syllables were produced in prosodically weak positions (Shen, 1990; Shih, 2005; Shih & Sproat, 1992; Xu, 1994, 1997). Syllables in these positions tend to be short and thus more susceptible to contextual influence. In contrast, the test syllables used in the current study were all utterance-final and prosodically prominent. Consequently, the test tones might be less likely to be influenced by the context. Indeed, our acoustic analyses from Experiment 1 indicated that these test syllables, even though read with the carrier phrases, were fairly long and that their \( f_0 \) contours closely resembled those produced in isolation.

Another factor that might have contributed to the limited amount of tonal coarticulation is the presence of a prosodic boundary before the test syllable. As noted, previous acoustic investigations on tonal coarticulation examined the second tone in a three-tone sequence, and the syllables in the sequence were in the same prosodic phrase. In Shen (1990), for example, speakers were instructed to read the non-word sequence with evenly distributed stress such that no single syllable was particularly prominent. In Xu (1994), the test syllable and the context formed a word or compound word. In contrast, in the current study, the coarticulation to be found would have taken place across word boundary. Thus, the speaker could have naturally inserted a prosodic boundary between the carrier phrase and the test syllable, limiting the degree of tonal coarticulation.

A third possibility is that in anticipation of the test tone, the speaker could have made comparable adjustments for both matching and mismatching the carrier-phrase-final tone. Consequently, no acoustic effect of tonal coarticulation could be observed on the test syllables. As reported earlier, however, our acoustic analyses did show statistically distinct onset \( f_0 \)'s for the test syllables when preceded by the two carrier tones. In other words, there was evidence of tonal coarticulation; the coarticulation was just not strong enough to make an impact on perception.

Nonetheless, some contrasts were still observed between this experiment and Experiment 1. In particular, Tone 1 identification improved because of reduced confusion with Tone 4. Similarly, Tone 2 identification also improved because of reduced confusion with Tone 3. These results suggest that the presence of the context still helped disambiguating these tone pairs. This is consistent with Gottfried and Suiter (1997), who found significantly less Tones 1–4 confusion in the onset-only syllables when the following context \( zi \) was available. In addition, although Tone 1 identification became as accurate as Tones 3 and 4, the RT measure showed that it was still not processed as quickly. Nonetheless, as was shown in Experiment 1, tone identification remained generally robust when the majority of the syllable was silenced.
5. Experiment 3

Experiment 2 showed that the presence of a context had some impact on tonal confusion pattern but only minimal effects on identification accuracy and RT. It was conjectured that the limited amount of tonal coarticulation in the stimuli might have contributed to the limited perceptual effects. In Experiment 3, the role of acoustic modification and phonetic context was further evaluated with a “cross-splicing” procedure. Recall that each test syllable used in Experiments 1 and 2 was recorded in two carrier phrases: one a match and the other a mismatch in \( f_0 \) between the offset of the carrier tone and the onset of the test tone. In this experiment, the test syllables were digitally removed from one carrier phrase and then appended with the other carrier phrase. The purpose was to evaluate whether the potential incompatibilities or discontinuities as a result of the cross-splicing would impact tone identification.

5.1. Method

The stimuli, participants, procedure, and data analysis in this experiment were identical to those used in Experiments 1 and 2 with the exception that the test syllables were now excised from the original carrier phrase and cross-spliced with the other carrier phrase. For example, the two versions of the syllable \textit{xing1} in \textit{Qing3 shuo1 xing1} (“Please say \textit{xing1}”) and \textit{Qing3 kan4 xing1} (“Please look at \textit{xing1}”) were excised from their carrier phrases, respectively. The \textit{xing1} originally from the \textit{Qing3 shuo1} ___ context was pasted onto the \textit{Qing3 kan4} ___ context. Similarly, the \textit{xing1} originally from the \textit{Qing3 kan4} ___ context was pasted onto the \textit{Qing3 shuo1} ___ context. That is, the final syllable of the carrier phrase either remained high-level or falling in \( f_0 \). The experiment was administered following a 5-min break after the participants completed Experiment 2.

5.2. Results

5.2.1. Accuracy

Fig. 7 shows tone identification accuracy by acoustic modification and tone. On average, identification accuracy was highest for center-only (97%) and intact syllables (97%), followed by silent-center syllables (93%), and lowest for onset-only syllables (83%). A repeated measures ANOVA revealed a main effect of modification: \( F(3, 117) = 97.21, p < .0001 \). All pair-wise means comparisons were significant \( (p < .0001) \) except for the difference between intact and center-only syllables. This pattern is identical to what was found in the
previous two experiments. That is, tone identification accuracy was generally a function of the amount of acoustic input available with the exception that center-only syllables were identified as accurately as intact syllables.

In addition, there was a main effect of tone: \( F(3, 117) = 14.501, p < .0001 \). On average, identification accuracy was higher for Tone 4 (96%), Tone 3 (95%), and Tone 1 (93%) than Tone 2 (85%). The three pairwise means comparisons involving Tone 2 were significant (\( p < .0001 \)). This pattern is identical to what was found in Experiment 2. Fig. 6 shows that the tone effect was not uniform across modifications, as indicated by the interaction between tone and modification: \( F(9, 351) = 16.29, p < .0001 \). The pattern of interaction was almost identical to the previous two experiments. That is, Tone 2 identification was compromised more than other tones in the silent-center condition.

Although no main effect of preceding tone was found, there was an interaction between tone and preceding tone: \( F(3, 117) = 15.97, p < .0001 \), which was not seen in previous two experiments. In particular, when the offset of the preceding tone was a match with the onset of the test tone, identification was more accurate for Tones 1 and 3 but less accurate for Tone 4. Recall that when the test syllables were presented in isolation (Experiment 1) or in their original carrier phrases (Experiment 2), the match or mismatch with a preceding tone did not affect response accuracy. That result was interpreted as the listeners’ use of syllable-intrinsic information as the primary source for tone identification. It was also speculated that the little impact the context had on tone identification probably arose from the limited amount of tonal coarticulation in the acoustic signal to begin with. That is, the canonical \( f_0 \) shape of a tone was largely preserved despite the presence of a preceding tone, which was supported by the acoustic analyses reported in Experiment 1. Here, however, we do see the impact of context, although only on selected tones. In particular, for Tones 1 and 3, the cross-spliced context disrupted identification less when the tones were originally produced in a matching context. Further, there was an interaction among tone, modification, and preceding tone: \( F(9, 351) = 11.68, p < .0001 \). While almost all tones in all modifications showed higher or comparable accuracy when the preceding tone was a match, Tone 2 in the silent-center condition and Tone 4 in the onset-only condition showed lower accuracy when preceded by a matching tone. It is not clear why these two conditions yielded the opposite and counterintuitive pattern. Nonetheless, these interaction effects involving the preceding tone clearly showed traces of tonal contextual effects on tone identification.

### 5.2.2. Confusion patterns

As in the previous two experiments, confusion matrices were generated for tone identification responses in the onset-only condition. Table 6 shows the results averaged across participants. A confusion matrix was generated for each participant; the numbers in the cells were then converted to percentages before being averaged across participants. The confusion pattern is very similar to what was found in Experiment 2 with minimal improvement in accuracy. The Tone 1–4 and Tone 2–3 confusions are still apparent, so is the bias towards Tones 3 in the Tone 2–3 confusion. In contrast, the asymmetry between Tones 1 and 4 seems to have disappeared. Tone 4 was no longer the preferred tone in the confusion. It is not clear why a cross-spliced context would help disambiguating Tones 1 and 4. It may be argued that the improvement arose from the participants’ familiarity with the stimuli and experimental task after participating in Experiments 1 and 2. However, if the practice effect were the only reason for the improvement, it should have reduced confusion across the board and not just for selected tones.

### 5.2.3. Reaction time

Fig. 7 shows tone identification RT by modification and tone. On average, RT was shortest for intact syllables (525 ms), followed by center-only (584 ms), silent-center (621 ms), and onset-only (654 ms) syllables. A repeated measures ANOVA revealed a significant main effect of modification: \( F(3, 117) = 19.72, p < .0001 \). All pair-wise means comparisons were significant except for the differences between center-only and silent-center, and between silent-center and onset-only syllables. Recall that in the previous two experiments there was virtually no RT difference among the three modified conditions. Now, it appears that the cross-spliced context took a toll on the speed of processing for the silent-center and onset-only syllables. When the test syllables were presented in isolation (Experiment 1) or with the original context (Experiment 2), the silent-center and onset-only syllables generated lower accuracy but comparable speed to the center-only syllables.
With the cross-spliced context (this experiment), the silent-center and onset-only syllables were responded to not only less accurately but also more slowly. The accuracy measure alone would not have revealed this pattern.

In addition, there was a main effect of tone: \(F(3, 117) = 21.07, p < .0001\). The RT was shorter for Tones 3 (476 ms) and 4 (540 ms) than for Tones 2 (673 ms) and 1 (695 ms). Pair-wise means comparisons showed significant differences between Tone 1 and Tones 3 and 4, as well as between Tone 2 and Tones 3 and 4. The tone effect was not uniform across modifications, as indicated by a significant interaction between tone and modification: \(F(9, 351) = 3.57, p < .0005\). Fig. 7 shows that among the modified conditions, RT was particularly long in the silent-center and onset-only syllables for Tones 1 and 2, suggesting that the identification of these two tones, compared with Tones 3 and 4, relied more heavily on syllable center and onset–offset combination.

Although no effect of the preceding tone was found, there was a significant interaction between tone and preceding tone: \(F(3, 117) = 4.09, p < .001\). In particular, when the preceding tone in the original stimuli was a match with the test tone, RT was shorter for Tones 1–3 but was longer for Tone 4. This result corroborates with the accuracy result discussed earlier, where cross-spliced Tones 1 and 3 were identified more accurately but Tone 4 less accurately with a matching preceding tone. Taken together, these results showed a fairly consistent pattern: cross-spliced Tones 1, 3, and perhaps 2 were facilitated by a matching preceding tone during recording, while cross-spliced Tone 4 was instead hindered by a matching preceding tone. Although the Tone 4 result was puzzling, these findings clearly indicate the impact of the tonal context.

5.2.4. Summary

The cross-splicing procedure revealed response patterns not seen in the previous two experiments. In particular, the impact of the preceding tone, which was not found when the test syllables were presented in isolation (Experiment 1) or in their original carriers (Experiment 2), showed up in several findings. The tone–preceding tone interaction and the tone–modification–preceding tone interaction indicated that for certain tones and certain modifications, identification accuracy and RT was indeed affected by the cross-spliced context. Specifically, for Tones 1 and 3, the cross-spliced context was less disruptive to tone identification accuracy when the tones were produced with a matching preceding tone. Similarly, for Tones 1–3, the cross-spliced context was less disruptive to identification latency when the tones were produced with a matching preceding tone. In contrast, for Tone 4, the cross-spliced context was more disruptive to identification accuracy and latency when the tone was produced with a matching preceding tone. While the reason for the Tone 4 result remains unclear, the results provided evidence that tonal context (i.e., the preceding tone) did influence tone identification.

What is the nature of the contextual influence? Recall from the acoustic analysis in Experiment 1 that the preceding tone had a small, but statistically significant acoustic effect on the mean \(f_0\) of the test tone averaged across the first six pitch periods. The modest carryover coarticulation was not sufficient to affect identification accuracy or RT when the test syllables were presented in isolation (Experiment 1) or in their original carrier phrases (Experiment 1). The null effect was taken to indicate the dominance of syllable-intrinsic information in tone identification, which is consistent with the acoustic analyses results. However, results from the cross-splicing manipulation indicated that tonal context did have an impact after all on both response accuracy and RT.

There were other differences between the current and previous experiments. Contrary to the previous two experiments where no RT difference was found among the modified conditions, the RT for the silent-center and onset-only syllables became significantly longer, indicating the cross-spliced context further impacted the speed of processing for these two types of modified syllables. In addition, the bias towards Tone 4 in the Tone 1–4 confusion has disappeared, although it is not clear why a cross-spliced context would facilitate the disambiguation.

5.2.5. All experiments considered

Now that tone identification from limited acoustic input has been evaluated in two preceding tone contexts and three splicing conditions, data from all three experiments were combined for a direct comparison. The average response accuracy was 87% (SD = 21) for isolated syllables, 90% (SD = 21) for syllables in original
carriers, and 92% (SD = 16) for cross-spliced syllables. The average RT was 758 ms (SD = 426) for isolated syllables, 648 ms (SD = 382) for syllables in original carriers, and 596 (SD = 301) for cross-spliced syllables. At first sight, the increase in accuracy and decrease in RT as the experiments progressed seems to indicate a practice effect. Since the experiments were administered in the same order for every participant with relatively brief breaks in between, the participants could have become familiar with the stimuli and/or the experimental task, thus facilitating their response. A repeated measures ANOVA on RT, with splicing, modification, tone, and preceding tone as fixed factors and participants as the random factor, did show a splicing effect: $F(2, 117) = 4.09, p < .05$. Pair-wise means comparisons indicated a significant difference between the excised (Experiment 1) and the cross-spliced (Experiment 3) conditions.

While practice could be a potential factor, particularly in response latency, it is not likely to be the only or major explanation for the response patterns observed in these experiments. First, an ANOVA on accuracy in fact showed no effect of splicing conditions. If there were a robust practice effect, there would have been a significant splicing effect in accuracy as well. Second, the stimuli were all high frequency, familiar items to the native listeners to begin with; thus familiarity is not likely to increase substantially for these participants. Task familiarity, on the other hand, was more likely to contribute to the potential practice effect, as was seen in the RT analysis. But even for the RT effect, only one out of three possible pair-wise comparisons turned out to be statistically reliable; therefore, there is no convincing evidence that the practice effect contributed significantly to the results found in this study. Nonetheless, it is acknowledged that for future studies, an ideal approach to within-subject design would be to randomize or vary the presentation order of the splicing conditions for each participant such that potential practice effects may be reduced or eliminated.

6. General discussion

In three experiments, the current study examined Mandarin tone identification by native listeners from partial acoustic input. By recording the test syllables in two carrier phrases and presenting the syllables in isolation, in original carriers, and in a cross-spliced context, this study also explored the effect of preceding context on tone identification. The use of the same signal processing procedure allowed a direct comparison with the results from Gottfried and Suiter (1997). Furthermore, the use of RT, in addition to response accuracy, provided a corroborating and more sensitive measure to real-time tone processing. Finally, more participants were used than those in Gottfried and Suiter (1997) to further enhance the reliability of the results.

Our perceptual data indicated that tone identification remained remarkably robust even when the majority of a syllable was missing and when the listeners were under time pressure to respond quickly. In particular, the silent-center syllables were identified with accuracy ranging from 86% (without context) to 93% (with context). These values are comparable to the accuracy of 91% (without context) and 95% (with context) reported by Gottfried and Suiter (1997). In addition, the accuracy values for the onset-only syllables (73% without context and 83% with context) were actually higher than those reported by Gottfried and Suiter (1997, 54% without context and 70% with context). With as few as six pitch periods or on average 17–39 ms of acoustic input, listeners in this study were able to achieve better-than-chance accuracy in tone identification, with or without context.

How do native listeners achieve such accurate tone identification with such limited acoustic input? For the center-only syllables, since the majority of the syllable was preserved, the majority of the $f_0$ contour was present. Not surprisingly, these syllables were identified with high accuracy. In fact, they were identified as accurately as the intact syllables. This result is consistent with Gottfried and Suiter (1997), who also showed comparable tone identification accuracy between the intact and center-only syllables (cf. Strange et al. (1983), who showed minimal increase in error rate for English vowel identification when the center of a vowel in a CV-consonant syllable was presented). However, our RT analyses in all three experiments revealed that the center-only syllables were not identified as fast as the intact syllables, with or without context. This indicated that the loss of onset and offset information, despite relatively minor effects on accuracy, still incurred a processing cost.

For the silent-center syllables, the majority of the syllable was silenced, thus the majority of the $f_0$ contour was not available. While one would expect substantially compromised identification performance because of
the missing $f_0$ information, Gottfried and Suiter (1997) showed that these syllables were in fact identified as accurately as the intact and center-only syllables (cf. Strange et al. (1983), who showed that silent-center vowels were identified as accurately as vowel center). In contrast, our accuracy and RT data jointly showed a different pattern. In particular, the silent-center syllables were identified less accurately than the intact and center-only syllables. The RT analysis further revealed that they were identified more slowly than the intact syllables but equally fast as the center-only syllables. Nonetheless, the overall accuracy was still remarkably high. In discussing the highly accurate vowel identification from silent-center English vowels, Strange et al. (1983) concluded that the dynamic spectral information in the transition from and to the adjacent consonants provided critical information for vowel identification. Analogously, the $f_0$ information available from the first six and final eight pitch periods also provided sufficient information for reliable tone identification. It appears that listeners were able to fill in the missing $f_0$ by using the available onset and offset information to infer the complete $f_0$ contour and to retrieve the tonal targets.

For the onset-only syllables, only the first six pitch periods remained in the acoustic signal. Since these syllables had the least amount of acoustic information, it was expected that they would be identified most poorly. That was indeed the finding from both Gottfried and Suiter (1997) and the current study. However, thanks to the use of the RT measure, the current study showed a different pattern for the identification of the onset-only syllables. Specifically, in Gottfried and Suiter (1997), the intact, center-only, and silent-center syllables were identified equally accurately and the onset-only syllables were identified more poorly. In contrast, Experiments 1 and 2 of the current study showed that identification accuracy decreased steadily as the amount of acoustic information became less, even though the RT among the modified syllables did not differ. In other words, when listeners were not under time constraints to make tone judgments (as in Gottfried & Suiter, 1997), the processing difference between the intact and center-only/silent-center syllables did not surface, i.e., all three were identified with comparable accuracy. When listeners were under time pressure to respond quickly (as in the current study), however, the accuracy difference emerged. Also, the non-significant decrement in RT for the modified syllables indicated no speed-accuracy tradeoff and thus validated the accuracy differences.

Despite these contrasts, both studies showed that the identification accuracy for these onset-only syllables was still well beyond chance. How was this achieved; especially, when these syllables were presented in isolation and the six pitch periods were all there were in the acoustic signal? Unlike the silent-center syllables, where listeners might be able to infer the complete $f_0$ contour by interpolation using the $f_0$ onset and offset, this strategy apparently would not work for the onset-only syllables because only the onset was available. As our acoustic analysis on the first six pitch periods showed, there was no discernible difference among the four tones in terms of $f_0$ shape, i.e., all appeared flat (Fig. 4); therefore, the $f_0$ contour shape is unlikely to be of use for tone identification. The acoustic analysis, however, did show statistically significant pair-wise differences in $f_0$ height among all four tones except for the contrast between Tones 1 and 4. It is clear from Fig. 4 that the four tones form two distinct groups: Tones 1 and 4 have a relatively high $f_0$ onset, and Tones 2 and 3 have a relatively low $f_0$ onset. Could it be that listeners managed to use the $f_0$ height information to identify the tones? However, since different speakers have different $f_0$ ranges, the judgment of high or low is most likely relative and would have to rely on the knowledge of the speaker’s $f_0$ range. There is some evidence that listeners engage in speaker normalization in tone perception. In particular, Leather (1983) and Moore and Jongman (1997) showed that the same $f_0$ contour could be identified as different tones depending on the perceived $f_0$ range of the speaker. Since only one speaker was used in the current study, our listeners could have obtained a pretty good sense of the speakers overall $f_0$ range, thus reducing the tone decision space. Assuming our listeners did manage to estimate the speaker’s $f_0$ range, an accuracy of 50% could be expected since the four Mandarin tones consist of two high-onset tones (Tones 1 and 4) and two low-onset tones (Tones 2 and 3). Our data indeed showed that excised Tone 2 without context was identified with 47% accuracy. However, the other three tones were identified with much higher accuracy ranging from 68% to 90%. Again, without distinct $f_0$ contour differences (Fig. 4), it is unlikely that $f_0$ shape could be of use. Although our acoustic analysis also indicates statistically significant duration difference among the tones (except between Tones 1 and 4), the range of the duration difference is rather small (17–39 ms, Table 3) and unlikely to be useful. Our speculation is that there could be other acoustic correlates to tonal distinctions (e.g., voice quality) that might be useful for tone identification in very brief segments. It has been noted that isolated Tone 3 for some speakers has a
likely due to the limited tonal coarticulation in the stimuli. In particular, compared with previous acoustic

discussed in the results and summary of that experiment, the absence of the preceding tone effect was most

was an interaction among tone, modification, and preceding tone for response accuracy in Experiment 2. As

(Experiment 1) or presented in original carriers (Experiment 2), the only effect involving the preceding tone

interactions involving the preceding tone were found. In particular, when the test syllables were excised

Experiment 2), the only effect involving the preceding tone was an interaction among tone, modification, and preceding tone for response accuracy in Experiment 2. As discussed in the results and summary of that experiment, the absence of the preceding tone effect was most likely due to the limited tonal coarticulation in the stimuli. In particular, compared with previous acoustic studies on tonal coarticulation (Shen, 1990; Shih, 2005; Shih & Sproat, 1992; Xu, 1994), the test syllables in the current studies were produced in a prosodically prominent position and the coarticulation would have been across words. Therefore, it was likely that our speaker inserted a prosodic boundary between the carrier and the test syllable, thus preventing or limiting the influence of the preceding tone. The acoustic analysis on the excised syllables indeed showed that the canonical $f_0$ shape was basically preserved, showing little deviation from the syllables produced in isolation.

Another potential reason for the lack of context effect is the nature of test syllables. As noted, 14 of the 24 syllables had a voiceless onset, which introduced a break in $f_0$ contour during the transition from the carrier to the test syllable. It could be that such a break reduced the impact of tonal coarticulation because the voiceless onset disrupted the continuity of the $f_0$ information. To evaluate this possibility, we conducted a number of post-hoc analyses comparing the identification accuracy and RT of syllables with voiced and voiceless onsets in all three experiments. It was expected that a syllable would be identified faster and more accurately if it had a voiced onset because voicing would be maintained and $f_0$ information would be available throughout. The results were inconclusive: For syllables presented in isolation (Experiment 1), those with a voiced onset were identified faster but there was no difference in accuracy. For syllables presented in their original carriers (Experiment 2), those with a voiced onset were identified with a higher accuracy but there was no difference in RT. No difference was found in accuracy or RT between voiced- and voiceless-onset syllables in the cross-spliced context (Experiment 3). This issue awaits further research.

Having said that, there were a number of occasions where the acoustic and perceptual impact of tonal coarticulation was observed. The acoustic analysis on the average $f_0$ of the first six pitch periods (reported in Experiment 1) showed a significant preceding tone effect, i.e., the average was higher when the test syllables followed the Tone 1 carrier as opposed to the Tone 4 carrier. There was no interaction between the test tone and the preceding tone, indicating that all four tones were comparably influenced by the preceding tone. Furthermore, when the context was cross-spliced (Experiment 3), identification of all but Tone 4 was facilitated in accuracy and/or RT when the tones were originally produced with a matching preceding tone. That is, the acoustic difference arising from the preceding tone was indeed noticed by the listeners. The perceptual effect did not show up when the test tones were presented in their original carriers, presumably because the listeners were able to compensate for the inherent tonal coarticulation generated in production (Xu, 1994). In contrast, when the test tones were presented in the more challenging, cross-spliced context, the effect of the preceding tone was able to emerge. In particular, when the test tone was originally produced with a matching preceding tone, it was easier for the canonical $f_0$ information to be preserved, thus minimizing the disruption from the cross-spliced context. In contrast, when the test tone was originally produced with a mismatching preceding tone, the test tone would be modified more by the carryover effect from the preceding tone, making test tone identification more challenging. These results are consistent with earlier studies showing the impact of tonal coarticulation on tone perception (Gottfried & Suiter, 1997; Lee, 2000; Xu, 1994).

Our data also revealed identification differences regarding specific tones. In general, Tones 3 and 4 were identified faster and more accurately than Tones 1 and 2. Tone 1 identification became as accurate as Tones 3 and 4 when the original carriers were present (Experiment 2), although the RT analysis revealed that it was still not processed as fast. Tone 2 identification was consistently the least accurate and slowest in all three experiments. These contrasts with the results from the pretest showing intact Tone 1 produced in isolation.
required the longest time to identify. It appears to suggest that the acoustic modifications introduced in this study was most detrimental to Tone 2 identification. This observation is further supported by the modification–tone interaction effects found in all three experiments (Figs. 5–7). In particular, Tone 2 identification was particularly compromised in the silent-center and onset-only syllables, indicating that Tone 2 identification relied on the syllable center more than other tones. Similarly, Tone 1 identification was also noticeably poorer in the onset-only syllables compared with Tones 3 and 4, suggesting that although Tone 1 identification was less susceptible to the missing syllable center, it could benefit from the presence of both the onset and offset.

The confusion pattern analyses on the onset-only syllables also revealed some error patterns commonly reported in the literature (i.e., Tones 2–3 confusion and Tones 1–4 confusion) and some asymmetries in the confusion patterns (i.e., Tone 2 misidentified as Tone 3 more often than vice versa; Tone 1 misidentified as Tone 4 more often than vice versa). The Tones 2–3 confusion and Tones 1–4 confusion are not surprising considering their onset f0 values. Our acoustic analysis showed that the four tones formed two distinct categories, with Tones 2 and 3 having a low f0 onset and Tones 1 and 4 having a high f0 onset. The preference for Tones 3 and 1, however, could not be predicted from the acoustic data. Since the onset-only syllables were only six pitch periods long and relatively flat (Fig. 4), it is unlikely that the asymmetries had their roots in the f0 contour shape. For the Tones 1–4 asymmetry, we speculated that upon hearing a high f0 onset, listeners had a tendency to give a Tone 4 response unless they had evidence of a flat contour for Tone 1. This interpretation is in fact consistent with the observation earlier that Tone 1 identification was particularly compromised when the center and offset was missing. For the Tones 2–3 asymmetry, it was conjectured that upon hearing a low f0 onset, listeners had a tendency to give a Tone 3 response unless they had evidence of a rising contour for Tone 2. This interpretation is also consistent with the observation earlier that Tone 2 identification was particularly compromised in the silent-center and onset-only condition. Specifically, previous acoustic studies and perceptual tests (Shen & Lin, 1991; Shen et al., 1993) showed that Tones 2 and 3, when produced in citation form, differ in the f0 “turning point”, or where the f0 contour turns from falling to rising. This turning point normally shows up in the syllable center and would be missing from the silent-center and onset-only syllables in the current study, which would explain why Tone 2 was identified particularly poorly from the these two types of syllables.

The finding that native listeners were able to identify incomplete tones with remarkable accuracy also has implications for the processing of lexical tone in spoken word recognition. In particular, the use of lexical tone in tone languages is for lexically contrastive functions; therefore the ultimate goal of processing tone is to associate sound with meaning. The listener’s ability to use limited acoustic information to derive tonal identity suggests that tones could actually be processed with high efficiency in lexical processing. Previous studies comparing the processing of lexical tone and segmental phonemes in speeded-response tasks showed longer RT and lower accuracy for tonal contrasts than segmental contrasts (Cutler & Chen, 1997; Ye & Connine, 1999). The interpretation has been that the order of information arrival in a given syllable during processing has to be consonant, vowel, and tone, since tonal information is assumed to be superimposed on the vowel and cannot be processed early. However, the finding from the current study would argue against the claim that the processing of tonal information occurs later than the processing of consonants and vowels. Indeed, there is evidence showing that lexical tone participates in early lexical processes as segmental structure does. For example, the disadvantage of tone processing could be neutralized with the presence of a semantic context (Ye & Connine, 1999). Tonal modifications could generate similar physiological response to semantic incongruity as segmental modifications (Schirmer, Tang, Penney, Gunter, & Chen, 2005). Tones were involved in lexical activation and competition (Lee, 2007). The results from the current study are consistent with these findings.

While the current study extended Gottfried and Suiter (1997) in demonstrating native listeners’ effective use of information from brief acoustic input in tone identification, certainly there are still unanswered questions awaiting further research. First, the use of fixed number of vocal pulses for stimuli preparation in both the current study and Gottfried and Suiter (1997) inevitably introduced variable stimulus durations, as was shown in the acoustic analyses. It will be worth investigating whether the use fixed durations would make a difference in the response pattern. In addition, native listeners’ ability to estimate f0 height could be implicated in their remarkable ability to identify the onset-only syllables. Although such estimation inevitably requires normalization for speakers, who can vary considerably due to anatomical, physiological, and other factors,
we speculated that voice quality could be a potential source of information. To our knowledge, there is no conclusive acoustic or perceptual evidence on the use of voice quality in Mandarin tone identification, let alone studies on very brief acoustic input. Future studies with more detailed acoustic analyses will be needed to address this possibility.

Acknowledgments

We thank Li Xu, Jessica Baker, Ann Marie Christy, Ning Zhou, and Danny R. Moates for assistance in data processing, and Deborah Siegrist, Yang Yang, Ning Zhou, and Aysun Balkan for assistance in administering the experiment. We also thank Terry Gottfried and two anonymous reviewers for their helpful comments on an earlier version of the paper.

References


