Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20). 2008. Volume 1.
Edited by Marjorie K.M. Chan and Hana Kang. Columbus, Ohio: The Ohio State University. Pages 147-164.

Talker and Contextual Effects On Identifying Fragmented Mandarin Tones

Chao-Yang Lee, Liang Tao and Z. S. Bond
Ohio University

This study investigated identification of fragmented Mandarin tones produced by single versus multiple speakers. Six minimal pairs, including all six Mandarin tonal contrasts, were digitally processed to generate intact, silent-center, center-only, and onset-only syllables. The syllables were produced either in isolation or with a carrier phrase *qing3 shuo3* __ ("Please say __"). The stimuli were presented in four blocks: (1) single speaker, isolated syllables; (2) single speaker, syllables with the carrier; (3) multiple speakers, isolated syllables; and (4) multiple speakers, syllables with the carrier. Forty native listeners and 55 non-native listeners were put under time pressure to identify the tones of the syllables and both response accuracy and reaction time were measured. Overall, the results showed higher accuracy for the single-speaker stimuli and when the syllables were presented with the carrier. For the native listeners, context facilitated identification of multiple-speaker stimuli more than single-speaker stimuli. For the non-native listeners, in contrast, context did not interact with the speaker effect. Identification of Tone 4 was consistently most accurate and least compromised by acoustic modification among the four tones. The results indicate different processing strategies for native and non-native listeners when dealing with incomplete acoustic input.

1. Introduction

Impoverished acoustic signal and inter-speaker variability are common challenges to speech perception. The acoustic signal to be deciphered by a listener is rarely clear and intact. The physical characteristics of a spoken message can also differ greatly among speakers despite identical linguistic content. Yet human listeners are known to be able to overcome these obstacles and uncover the linguistic representation intended by the speakers. Research on speech perception has identified many sources of information available in the acoustic signal and the phonetic knowledge involved in the decoding process (Pisoni & Remez, 2005). Cross-linguistic studies have further revealed similarities and differences between native and non-native perception of phonological contrasts (Sebastián-Gallés, 2005). Together they suggest a highly efficient speech perception system conditioned by the nature of the listener's linguistic background.

The purpose of this study was to examine how native and non-native listeners identify Mandarin tones from incomplete acoustic input and how the two groups of listeners deal with speaker variability. To these ends, we adopted the "silent-center" paradigm (Strange, Jenkins, & Johnson, 1983), where various parts of a syllable were digitally silenced, leaving only partial input available. Identification of tones from these fragmented syllables was compared between syllables produced by a single speaker and those produced by multiple speakers. Comparisons were also made between tones produced in isolation and those produced with a carrier phrase to evaluate the contribution of phonetic context. Finally, participants were put under time pressure to make tone judgments. Reaction time was measured in addition to the traditional accuracy measure to explore the online nature of lexical tone processing.

The silent-center paradigm has provided important evidence for how vowels and lexical tones are identified from incomplete acoustic signal. In particular, studies on English vowels have shown that listeners are capable of using dynamic spectral information from consonant-vowel and vowel-consonant transitions to identify vowels despite missing steady-state formant information (e.g., Strange, Jenkins, & Johnson, 1983). Analogously, listeners of Mandarin Chinese have been shown to be able to identify silent-center tones as accurately as intact and center-only tones (Gottfried & Suiter, 1997), indicating the perceptual system's ability to integrate information from syllable onset and offset for tone identification. Using a speeded-response version of the task, Lee, Tao, and Bond (2008) showed that silent-center tones were not identified as accurately as intact and center-only tones under time pressure. The reaction time analyses further revealed that center-only tones were not identified as quickly as intact tones despite comparable accuracy. Despite the processing cost incurred under time pressure, identification of these fragmented tones remained highly accurate. In sum, these studies showed that the center of the syllable, which was traditionally regarded as providing the critical information, is not necessary for reliable identification of vowels or tones.

Furthermore, silent-center studies on lexical tones have also revealed processing differences between native and non-native listeners in the use of context for tone identification. Gottfried and Suiter (1997) found that the addition of a syllable following the test tone significantly facilitated tone identification accuracy for native but not non-native listeners. Lee, Tao, and Bond (submitted) showed that non-native listeners were not compromised as native listeners were (Lee et al., 2008) by test tones that had been edited out from a precursor carrier phrase and "cross-spliced" with a different carrier phrase. Earlier studies focusing on the perceptual impact of tonal coarticulation had also shown that native listeners were sensitive to contextual tonal variation (Xu, 1994). Taken together, these findings suggest that native tone identification is characterized by efficient use of tonal context, while non-native listeners focus on syllable-intrinsic information and make relatively little use of contextual tonal information for tone identification. Specifically, since non-native listeners did not pay much attention to context, adding a context did not help (Gottfried & Suiter, 1997). Presenting a conflicting context did not hurt their tone identification either (Lee et al., submitted).

Incorporating speaker variability into this line of inquiry allows further evaluation of fragmented tone identification by native vs. non-native listeners in important ways. How listeners deal with speaker variability has been a central issue in phonetic perception research (Johnson, 2005). As noted, speakers differ in vocal tract characteristics, thus the acoustic signal generated by different speakers will inevitably show variability even when the linguistic content is the same. Numerous studies have shown that listeners take into consideration context as a reference frame for phonetic perception (e.g., Ladefoged & Broadbent, 1957). Speaker variability is certainly an issue for lexical tone perception, which relies primarily on the perception of a speaker's f_0. Since f_0 range varies across speakers, the actual f_0 for a given tone produced by different speakers will most likely show variability as well. Not surprisingly, native listeners take into consideration the information provided by context in tone identification (Leather, 1983; Lin & Wang, 1985; Fox & Qi, 1990; Moore & Jongman, 1997; Wong & Diehl, 2003). These findings indicated that listeners engage in some kind of speaker normalization process for tone perception just as they do for vowel perception.

What is the role of speaker variability in perceiving acoustically incomplete tones? How would context impact identification of tones produced by multiple speakers? Will the processing difference found earlier between native and non-native listeners manifest itself in the same way when speaker variability is introduced? These issues are involved in explicating the perceptual system's ability to compensate for impoverished acoustic signal. These issues are also important in accounting for the nature of native vs. non-native processing of lexical tones hinted in earlier studies (Gottfried & Suiter, 1997; Lee et al., 2008; Lee et al., submitted).

Predictions can be made regarding the potential effects of speaker variability, context, and linguistic experience. First, it is conceivable that speaker variability adds to the processing demand; therefore tone identification from a multiple-speaker stimulus set should be more error-prone and time-consuming than from a single-speaker set. Second, context offers more information about individual speaker characteristics and should facilitate tone identification from a multiple-speaker stimulus set more than a single-speaker set. Finally, speaker variability might pose a greater challenge to non-native listeners; therefore their identification performance would be compromised more than native listeners.

In sum, the objective of this study was to answer the following questions: Do listeners identify tones produced by multiple speakers at the same level as tones produced by one speaker? To what extent does linguistic context affect the identification of the tones? Do native and non-native listeners use different strategies when dealing with Mandarin multiple-speaker tones with and without context? How do native and non-native listeners respond to tone fragments representing different sources of information? What are the acoustic bases for tone judgments? Below we report a perception experiment where native and non-native listeners attempted tone identification from syllables that were digitally processed to generate four types of stimuli varying in the amount of acoustic information. Results from acoustic analyses were then presented to discuss the acoustic bases for the tone identification performance.

2. Perception Experiment

2.1 Method

2.1.1 Materials

Six minimal tone pairs were selected including all six tonal contrasts in Mandarin: 1-2 (*xing* 星行), 1-3 (*xi* 西洗), 1-4 (*si* 司四), 2-3 (*hai* 还海), 2-4 (*shi* 十是), and 3-4 (*da* 打大). All of these are high-frequency, common words known by all participants. All participants were also familiar with the convention of designating Mandarin tones by the numbers 1, 2, 3, and 4.

Ideally, minimal tone pairs should be selected from the same set of syllables to control for segmental structure, considering the potential effects of syllabic structure on tone (Shih, 1987). However, since the same set of stimuli would be presented to non-native listeners, our primary concern was that the selected words should have been learned and known by the non-native learners as well. This decision was also motivated by many findings that lexical status and word frequency/familiarity can impact phonetic perception (Ganong, 1980; McQueen, 1991; Connine, Clifton, & Cutler, 1987), including tone identification (e.g., Fox & Unkefer, 1985). Nonetheless, efforts were made to select words with simple syllabic structure. In the end, all selected words were consonant-vowel or consonant-vowel-nasal syllables.

To examine the effect of context, the 12 syllables were read in isolation and with a carrier phrase *Qing3 shuo1* __ (“Please say __”). To examine the effect of speaker, the syllables were recorded by five native speakers of Mandarin. One female speaker (from Shandong) was used for the single-speaker conditions; two additional female speakers (one from Beijing and one from Changchun) and two male speakers (both from Beijing) were used for the multiple-speaker conditions.

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 test syllable was digitally modified with BLISS to generate four types of syllables: intact, center-only, silent-center, and onset-only. In particular, the first six and final eight pitch periods of the intact syllables were digitally edited to generate the modified syllables. The center-only syllables were constructed by removing the first six and final eight pitch periods of a syllable. The silent-center syllables were generated by preserving only the first six and final eight pitch periods. The onset-only syllables were produced by preserving only the first six pitch periods of the syllables. 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. A total of 480 stimuli (4 tones × 3 tokens per tone × 4 modifications × 2 contexts × 5 speakers) were used in this experiment.

2.1.2 Participants

The native listeners included 40 native speakers of Mandarin recruited from the Ohio University community with cash compensation. They included 20 females (mean age = 27, SD = 4.5) and 20 males (mean age = 26, SD = 4.7). All spoke Mandarin on a daily basis and none reported any speech or hearing difficulties. Twenty-five participants reported speaking some dialect of Chinese other than Mandarin, but all identified Mandarin as their native language.

The non-native listeners included 55 Chinese language students at Ohio University. The participants included 33 first-year (14 female & 19 male) students, 16 second-year (five male & 11 female) students, and six third-year (three female & three male) students. At the time of testing, the first-year, second-year, and third-year students had taken approximately three, six, and nine academic quarters of Chinese language classes. Ideally, the number of participants would be evenly distributed across levels of instruction and experience. For this study, we tested all available members of the target population. The non-native participants received partial course credit for participating in this study.

2.1.3 Procedure

The stimuli, saved as individual audio files, were imported to AVRunner, the subject-testing program in BLISS, for stimulus presentation. The 480 items were divided into four blocks: single-speaker, presented in isolation (48 items); single-speaker, presented with a carrier phrase (48 items); multiple-speaker, presented in isolation (192 items); and multiple-speaker, presented with the carrier phrase (192 items). For each participant, AVRunner assigned a uniquely randomized presentation order such that no two participants received the same order of presentation. The order of presentation for the blocks was also randomized.

Participants were tested individually in a quiet room in the Department of Linguistics at Ohio University. They listened to the stimuli through a pair of high-quality headphones (Koss R80) connected to a Toshiba laptop computer. The participants were instructed to identify the tone of each syllable by pressing buttons labeled "1", "2", "3", and "4" on the computer keyboard, representing the four Mandarin tones. They were told that some of the syllables have been digitally processed such that parts of the syllable might be missing. They were instructed to make the best guess of the stimulus tones and to respond as quickly as possible.

2.1.4 Data analysis

Response accuracy and reaction time were recorded by BLISS automatically. Reaction time was measured from stimulus offset to avoid the potential confound of intrinsic duration differences among the tones. Only correct responses were included in the reaction time analysis.

For the native listeners, response data from all speaker and context conditions were first combined to evaluate the effects of speaker and context. Responses were then analyzed separately for the four blocks of stimuli to evaluate the effect of acoustic modification and tone under the four speaker (single vs. multiple) and context (isolated vs. contextual) conditions. For each block, repeated measures ANOVAs were conducted on response accuracy and reaction time with acoustic modification (intact, center-only, silent-center, & onset-only) and stimulus tone (1, 2, 3, & 4) as fixed factors and participants as a random factor. 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%.

For the non-native listeners, response data from all speaker and context conditions were also combined to evaluate the effects of speaker, context, and year of Chinese instruction. Responses were then analyzed separately for the four blocks of stimuli to evaluate the effect of acoustic modification and tone under the four speaker (single vs. multiple) and context (isolated vs. contextual) conditions. As in the native data analysis, for each block, ANOVAs were conducted on response accuracy and reaction time with acoustic modification (intact, center-only, silent-center, & onset-only) and stimulus tone (1, 2, 3, & 4) as within-subject factors, year of Chinese instruction (first-year, second-year, & third-year) as a between-subject factor, and participants as a random factor. 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%.

2.2 Results

2.2.1 Effects of speaker and context

Table 1 shows the average accuracy of tone identification by speaker and context. For the native data, ANOVAs revealed significant main effects of speaker ($\underline{F}$ (1, 39) = 59.98, $\underline{p}$ < .0001) and context ($\underline{F}$ (1, 39) = 119.05, $\underline{p}$ < .0001), and a significant speaker-context interaction ($\underline{F}$ (1, 39) = 21.32, $\underline{p}$ < .0001). Overall, accuracy was higher for the single-speaker stimuli (89%) than for the multiple-speaker stimuli (86%). Accuracy was also higher for tones presented in context (90%) than for tones presented in isolation (85%). The context effect, however, was not uniform across single- and multiple-speakers. In particular, context facilitated tone identification more when listeners heard the multiple-speaker stimuli than when they heard single-speaker stimuli.

For the non-native data, ANOVAs revealed significant main effects of speaker ($\underline{F}$ (1, 52) = 18.05, $\underline{p}$ < .0001) and context ($\underline{F}$ (1, 52) = 15.21, $\underline{p}$ < .0005). As in the native data, accuracy was higher for single-speaker stimuli (65%) than for multiple-speaker stimuli (62%). Accuracy was also higher for tones presented in context (65%) than for tones presented in isolation (62%). Unlike the native listener data, the speaker-context interaction only approached significance (F (1, 52) = 3.76, p = .058), indicating that the effect of context was uniform across single- and multiple-speaker stimuli. Although third-year students (73%) clearly outperformed first-year (63%) and second-year (61%) students, the difference was not statistically significant.

Table 1 ***Average accuracy (in percentage) of tone identification by speaker, context, and linguistic background. Standard deviation is shown in parenthesis.***

	Speaker	Isolated tones	Contextual tones
First-year	Single	64 (34)	66 (35)
	Multiple	59 (29)	63 (28)
Second-year	Single	60 (34)	65 (32)
	Multiple	57 (29)	61 (28)
Third-year	Single	76 (32)	74 (34)
	Multiple	68 (26)	73 (25)
Native	Single	88 (22)	91 (20)
	Multiple	82 (22)	90 (15)

Finally, since the assignment of speakers into the single- and multiple-speaker conditions was arbitrary, it was necessary to ensure that the speakers were equally intelligible to the listeners in the first place. To this end, the native listeners' accuracy of response to intact syllables produced by the five speakers was taken as the index of intelligibility and was further analyzed. A repeated measures ANOVA was conducted on response accuracy with speaker (five speakers) and context (isolated & contextual) as fixed factors and participants as a random factor. The ANOVA revealed no effect of speaker (F (4, 156) = 0.83, p = .51), a significant effect of context (F (1, 39) = 7.22, p < .05), and a significant speaker-context interaction (F (4, 156) = 2.69, p < .05). The lack of speaker effect indicates that all five speakers were equally intelligible; therefore any effect found in the single- vs. multiple-speaker comparison would not be due to speaker intelligibility issues. Consistent with earlier results, response to intact contextual tones (96%) was more accurate than isolated tones (94%) for all participants. Inspection of the speaker-context interaction revealed that the interaction arose from the greater improvement for a male speaker compared to other speakers when context was added. Nonetheless, the overall null effect of speaker indicated that all speakers were equally intelligible.

2.2.2 Single-speaker, tones in isolation

For the native data, ANOVAs revealed significant main effects of modification (F (3, 117) = 106.82, p < .0001) and tone (F (3, 117) = 11.53, p < .0001), and a significant modification-tone interaction (F (9, 351) = 7.75, p < .0001). As expected, intact syllables (97%) and center-only syllables (98%) were identified most accurately, followed by silent-center syllables (85%) and onset-only syllables (70%). Pair-wise means comparisons showed all contrasts were significantly different except between intact and center-only syllables. Tone 4 was identified more accurately (95%) than Tone 3 (87%), Tone 1 (85%), and Tone 2 (83%). Pair-wise means comparisons showed all contrasts involving Tone 4 were significant. The interaction shows that for Tones 1, 2, and 3, accuracy dropped significantly for the silent-center and onset-only syllables. In contrast, the modifications hardly compromised Tone 4 identification.

Just as for the native listener data, for the non-native listener data, ANOVAs revealed significant main effects of modification (F (3, 156) = 70.04, p < .0001) and tone (F (3, 156) = 18.74, p < .0001), and a significant modification-tone interaction (F (9, 468) = 7.75, p < .0001). The main effect of year of instruction only approached significance (F (2, 52) = 3.06, p = .056).

The difficulty of identifying syllable modifications followed the native listener patterns. Intact syllables (77%) and center-only syllables (76%) were identified most accurately, followed by silent-center syllables (59%) and onset-only syllables (45%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables. Tone 4 (77%) and Tone 1 (69%) were identified more accurately than Tone 2 (57%) and Tone 3 (53%). Pair-wise means comparisons showed all contrasts between the two groups were significant. Third-year students (76%) outperformed first-year (63%) and second-year (60%) students, although these differences only approached significance.

The interaction shows that acoustic modification influenced identification of the four tones in quite different ways. Specifically, removing the onset and offset did not influence Tone 1 identification; Tone 2 actually benefited from highlighting the center information; Tone 3 identification accuracy dropped linearly as less acoustic input became available; and Tone 4 was least compromised by acoustic modification.

2.2.3 Single-speaker, tones in context

For native listeners, ANOVAs revealed significant main effects of modification (F (3, 117) = 81.02, p < .0001) and tone (F (3, 117) = 11.49, p < .0001), and a significant modification-tone interaction (F (9, 351) = 18.28, p < .0001). Intact syllables (98%) and center-only syllables (97%) were identified most accurately, followed by silent-center syllables (93%) and onset-only syllables (76%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables and between center-only and silent-center syllables. Tone 4 (96%) and Tone 3 (94%) were identified more accurately than Tone 2 (89%) and Tone 1 (85%). Pair-wise means comparisons showed the two groups are different from each other except between Tones 3 and 2. The interaction shows that identification of silent-center tone identification improved significantly with the addition of context and that Tone 4 remains quite accurate despite acoustic modification.

For the non-native listeners, ANOVAs revealed significant main effects of modification (F (3, 156) = 56.65, p < .0001) and tone (F (3, 156) = 16.9, p < .0001), and a significant modification-tone interaction (F (9, 468) = 9.46, p < .0001). Intact syllables and center-only syllables (both at 78%) were identified most accurately, followed by silent-center syllables (59%) and onset-only syllables (49%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables. Tone 4 (80%) was identified most accurately, followed by Tone 1 (68%) and Tone 3 (63%); Tone 2 (53%) was identified least accurately. All pair-wise means comparisons were significant except for the contrast between Tones 1 and 3. The interaction shows virtually the same pattern as the isolated tones reported earlier, indicating the addition of context did not change the pattern of response substantially for the non-native listeners.

2.2.4 Multiple-speaker, tones in isolation

For the native data, ANOVAs revealed significant main effects of modification (F (3, 117) = 385.04, p < .0001) and tone (F (3, 117) = 17.17, p < .0001), and a significant modification-tone interaction (F (9, 351) = 26.11, p < .0001). Intact syllables (97%) and center-only syllables (94%) were identified most accurately, followed by silent-center syllables (76%) and onset-only syllables (61%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables. Tone 4 (89%) was identified more accurately than Tone 3 (82%) and Tone 1 (82%); and Tone 2 (75%) was identified least accurately. All pair-wise means comparisons were significant except between Tones 3 and 1. The interaction shows that center-only syllables were identified as accurately as intact syllables for all tones but Tone 3. Identification of silent-center and onset-only tones in contrast were compromised across the board, even for Tone 4. This indicates that talker variability added difficulty to tone identification particularly in the silent-center and onset-only syllables.

For the non-native data, ANOVAs revealed significant main effects of modification (F (3, 156) = 179.46, p < .0001) and tone (F (3, 156) = 25.45, p < .0001), and a modification-tone interaction (F (9, 468) = 15.52, p < .0001). Intact syllables (77%) were identified most accurately, followed by center-only syllables (71%), silent-center syllables (50%) and onset-only syllables (38%). Pair-wise means comparisons showed all contrasts were significant. Tone 4 (73%) was identified most accurately, followed by Tone 1 (59%) and Tone 2 (56%), and Tone 3 (48%) was identified least accurately. All pair-wise means comparisons were significant except between Tones 1 and 2. The interaction shows that the impact of modification varied across tones, similar to the previous two conditions. The addition of more speakers did not seem to change the response pattern of the non-native listeners.

2.2.5 Multiple-speaker, tones in context

For the native data, ANOVAs revealed significant main effects of modification (F (3, 117) = 196.79, p < .0001) and tone (F (3, 117) = 17.04, p < .0001), and a significant modification-tone interaction (F (9, 351) = 18.8, p < .0001). Intact syllables (98%) and center-only syllables (97%) were identified most accurately, followed by silent-center syllables (87%) and onset-only syllables (78%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables. Tone 4 (95%), Tone 3 (91%), and Tone 1 (90%) were identified more accurately than Tone 2 (84%). All pair-wise means comparisons were significant except between Tones 4 and 3 and between Tones 3 and 1. The interaction shows that the addition of context in the multiple-speaker setting appeared to have facilitated identification of onset-only tones most substantially. This is conceivable. In particular, since onset-only tones were deprived of the majority of syllable-intrinsic tonal information, the addition of the context could have provided extrinsic information that would be particularly useful when speaker variability exists.

For the non-native data, ANOVAs revealed significant main effects of modification (F (3, 156) = 147.77, p < .0001) and tone (F (3, 156) = 24.84, p < .0001), and a significant modification-tone interaction (F (9, 468) = 14.55, p < .0001). Intact syllables (77%) and center-only syllables (74%) were identified most accurately, followed by silent-center syllables (57%) and onset-only syllables (48%). Pair-wise means comparisons showed all contrasts were significant except between intact and center-only syllables. Tone 4 (80%) was identified more accurately than Tone 1 (64%), Tone 3 (57%), and Tone 2 (55%). All pair-wise means comparisons were significant except between Tones 1 and 3 and between Tones 3 and 2. The interaction plot shows a similar pattern to the isolated condition.

3. Acoustic Analyses

The perception experiment showed that for both groups, identification was more accurate for single-speaker tones and when the tones were presented in context. However, compared to non-native listeners, native listeners were facilitated more by context when dealing with multiple-speaker stimuli. Detailed analyses of the four blocks also revealed that native listeners' response to modified tones were influenced by speaker and context, but non-native listeners' response pattern remained quite consistent irrespective of speaker or context. One possibility is that native listeners were more sensitive to changes in the acoustic signal resulting from the speaker and context variations. But what are the specific acoustic changes involved? To explore the acoustic basis of the perceptual response patterns, acoustic analyses were conducted on duration and fundamental frequency of the stimuli, the two acoustic measures that are most relevant to tone identification.

3.1 Results

3.1.1 Duration

Table 2 shows the average duration of three components of the syllable rhyme, which carries f_0 information: the first six pitch periods, the center, and the final eight pitch periods. ANOVAs were conducted with speaker (single & multiple), context (isolated & contextual), and tone (1, 2, 3, & 4) as fixed factors on the duration of the three components. 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%.

Table 2 ***Average duration (in ms) of the three components of the syllable rhyme that carries f_0 information. Standard deviation is shown in parenthesis.***

Isolated tones	Tone	First six periods	Center	Final eight periods
Single	1	19 (0)	307 (23)	26 (1)
	2	28 (2)	363 (45)	24 (2)
	3	31 (1)	447 (53)	39 (1)
	4	18 (1)	241 (48)	68 (20)
Multiple	1	27 (8)	326 (81)	36 (11)
	2	35 (10)	364 (75)	40 (12)
	3	38 (13)	381 (99)	57 (12)
	4	26 (7)	211 (77)	64 (12)

Contextual tones	Tone	First six periods	Center	Final eight periods
Single	1	17 (1)	269 (2)	25 (1)
	2	29 (0)	349 (40)	26 (0)
	3	32 (1)	404 (9)	40 (1)
	4	18 (3)	190 (16)	80 (25)
Multiple	1	26 (7)	264 (68)	35 (10)
	2	36 (10)	286 (67)	40 (13)
	3	38 (12)	311 (105)	57 (14)
	4	25 (6)	136 (59)	92 (35)

For the duration of the first six periods, the ANOVA showed significant main effects of speaker (F (1, 104) = 13.78, p < .0005) and tone (F (3, 104) = 11.78, p < .0001). The average duration of syllable onsets was longer for the multiple-speaker syllable fragments (31 ms) than for the single-speaker syllable fragments (24 ms). This difference was expected, as the multiple-speaker stimuli included items produced by two male speakers, whose longer period contributed to the higher average. This observation was confirmed by Table 3, which lists the average duration of the first six pitch periods for individual speakers. In addition, the average duration for Tone 2 (34 ms) and Tone 3 (37 ms) was also longer than for Tone 1 (25 ms) and Tone 4 (24 ms). This finding was also expected since the former set of tones starts with lower f_0 values, implying longer periods.

Table 3 ***Average duration (in ms) of the first six pitch periods, center, and final eight pitch periods of the syllable rhyme for individual speakers. Standard deviation is shown in parenthesis.***

Condition	Speaker	First six periods	Center	Final eight periods
Single	Female	24 (6)	321 (87)	41 (23)
Multiple	Female 1	24 (3)	362 (85)	50 (38)
	Female 2	24 (5)	321 (106)	44 (19)
	Male 1	34 (7)	183 (69)	51 (12)
	Male 2	44 (9)	275 (84)	65 (16)

For the duration of the center, the ANOVA showed significant main effects of speaker (F (1, 104) = 4.45, p < .05), tone (F (3, 104) = 22.86, p < .0001), and context (F (1, 104) = 9.89, p < .005). The average duration of syllable centers was longer for single-speaker stimuli (321 ms) than multiple-speaker stimuli (285 ms). The average duration for Tone 3 (362 ms), Tone 2 (331 ms), and Tone 1 (294 ms) was longer than Tone 4 (182 ms), attributable to inherent differences in f_0. All pair-wise comparisons were significant except between Tones 3 and 2 and between Tones 2 and 1. Tones on isolated syllables (324 ms) were longer than tones in context (260 ms), which was conceivable.

For the duration of the final eight periods, the ANOVA showed significant main effects of speaker (F (1, 104) = 9.85, p < .005) and tone (F (3, 104) = 32.76, p < .0001). The average duration was longer for multiple-speaker stimuli (53 ms) than single-speaker stimuli (41 ms). Again, this finding was to be expected, as the multiple-speaker stimuli included items produced by two male speakers, whose longer period contributed to the higher average. Again, this observation was confirmed by Table 3, which also lists the average duration of the final eight pitch periods for individual speakers. The average duration for Tone 4 (77 ms) was longer than Tone 3 (53 ms), which is in turn longer than Tone 2 (37 ms) and Tone 1 (34 ms). All pair-wise means comparisons were significant except between Tones 2 and 1. Again, this was expected since Tone 4 ends with a lower f_0 and thus longer periods than the other tones.

3.1.2 Fundamental frequency

The f_0 contours of the intact syllables by speaker and context were generated by BLISS with autocorrelation. The shapes of these contours are consistent with traditional descriptions of the four Mandarin tones. No discernible f_0 differences were observed between the isolated and contextual tones.

To obtain a more fine-grained measure of the f_0 information available in the partial acoustic input, the f_0 of the first six pitch periods for each syllable was measured. This was accomplished by manually marking each of the six glottal cycles on the waveform display. Overall, for all five speakers, the f_0 contours of the first six pitch periods were basically flat for all four tones. In addition, the four tones formed two distinct sets, with Tones 1 and 4 having higher f_0 values and Tones 2 and 3 having f_0 lower values. The speaker used in the single-speaker condition (Speaker 1) appeared to have greater separation between the two sets of tones than the speakers used in the multiple-speaker condition (Speakers 2, 3, 4, & 5). Again, there were no discernible differences between isolated and contextual tones.

To evaluate these observations quantitatively, a mean f_0 was obtained by averaging across the six f_0 values for each tone. Since the f_0 contours were all flat and similar to each other, the mean f_0 should be a reasonable summary measure. An ANOVA was conducted on the mean f_0 with speaker (single & multiple), context (isolated & contextual), and tone (1, 2, 3, & 4) as fixed factors. The ANOVA showed significant main effects of speaker ($\underline{F}$ (1, 104) = 22.61, $\underline{p} < .0001$) and tone ($\underline{F}$ (3, 104) = 27.5, $\underline{p} < .0001$), and a significant speaker-tone interaction ($\underline{F}$ (3, 104) = 2.97, $\underline{p} < .05$). It is not surprising that the single-speaker stimuli (269 Hz) had a higher average f_0 than the multiple-speaker stimuli (212 Hz) because latter set included two male speakers. Tone 1 (261 Hz) and Tone 4 (272 Hz) had higher mean f_0 values than Tone 2 (186 Hz) and Tone 3 (174 Hz). Post-hoc tests showed that all pair-wise comparisons between the former and latter sets of tones were significant, confirming an earlier observation that the four tones form two distinct sets in onset f_0. The speaker-tone interaction showed that the difference between the two sets of tones was greater in the single-speaker condition, again confirming an earlier observation.

4. General Discussion

The research questions in this study dealt with the effect of speaker variability on the identification of incomplete Mandarin tones, the contribution of context to the identification, and the response differences between native and non-native listeners. A perception experiment was conducted to evaluate the accuracy and reaction time of responses to fragmented tones in various speaker-context conditions, and acoustic analyses were conducted to evaluate the bases of those responses.

Speaker variability clearly added processing demand to tone identification. For all listeners, tone identification was less accurate for multiple-speaker stimuli. For non-native listeners, tone identification was also slower for multiple-speaker stimuli. The acoustic analyses showed that the duration and f_0 of the tones were indeed much more variable in the multiple-speaker stimuli than in the single-speaker stimuli, providing partial explanation for increased difficulty of tone identification when listening to multiple talkers.

The presence of context—a short precursor carrier phrase—also facilitated tone identification. For all listeners, tone identification with context was more accurate. For the native listeners, identification of tones presented in context was also faster. More interestingly, for the native listeners, context was particularly helpful for the multiple-speaker stimuli, as shown in both accuracy and reaction time. In contrast, for the non-native listeners, context was equally helpful irrespective of single-speaker or multiple-speaker stimuli. In other words, when dealing with multiple-speaker stimuli, native listeners benefitted from the presence of context more than the non-native listeners did.

There are two possible ways context could have facilitated tone identification. One is the listeners' knowledge of tonal coarticulation, which has been shown to impact tone perception by native listeners (Xu, 1994; Gottfried & Suiter, 1997; Lee et al., 2008). The other possibility is that context provides information for speaker normalization (Leather, 1983; Lin & Wang, 1985; Fox & Qi, 1990; Moore & Jongman, 1997; Wong & Diehl, 2003). While both accounts are consistent with current findings that all listeners showed improvement with context, our acoustic analyses did not find evidence for the existence of tonal coarticulation in the tone stimuli presented in context. As noted, the absence of acoustic influence from the precursor carrier tone was most likely due to the test tone being in a prosodically strong position, which prevented the anticipatory coarticulation from occurring. Nonetheless, tones presented in context were still identified more accurately, suggesting that the primary use of context was to establish a reference frame for speaker normalization.

This interpretation is also corroborated by the native data, which showed identification of silent-center and onset-only tones improved most dramatically with the addition of context. In particular, for single-speaker stimuli, the addition of context greatly increased accuracy for silent-center tones. For multiple-speaker stimuli, the addition of context greatly increased accuracy for onset-only tones. Recall that silent-center and onset-only tones were deprived of the majority of the f_0 contour in the middle of a syllable, thus the direction of f_0 movement was not physically present. It is conceivable that context offers a reference frame by exposing the speaker's f_0 range; therefore the extrinsic information could be used to infer tone identity when syllable-intrinsic f_0 information was largely unavailable. Note though that this interpretation applies only to the native data. The non-native data did not show distinct response patterns between isolated and contextual tones for either single-speaker stimuli or multiple-speaker stimuli.

The similarities and differences between the native and non-native data should now become obvious throughout. Both groups of listeners identified tones with lower accuracy when the tones were produced by multiple speakers and when the tones were presented in isolation. As noted, these results are not surprising given the known facilitative effect of context and the challenge of adapting to multiple speaker voices. The differences between the two groups of listeners showed up in two major ways. First, native listeners were facilitated by context more when dealing with multiple speakers than a single speaker; non-native listeners did not show such a preference. Second, for native listeners, identification of silent-center and onset-only tones improved greatly with the presence of context; non-native listeners did not show this preference.

Together these two findings suggest that native listeners were more sensitive to information extrinsic to the test syllable for tone identification, while the non-native listeners relied primarily on syllable-intrinsic information. In particular, as the acoustic analyses showed, the multiple-speaker set introduces more acoustic variability to the duration and f_0 of the tones. The lack of tonal articulation in the signal also indicated that context was used primarily to establish a reference frame for f_0 judgments in the multiple-speaker set. Since the non-native listeners did not benefit from context in the multiple-speaker set any more than in the single-speaker set, it suggests that the non-native listeners were not relying on speaker normalization as much to aid tone identification. Furthermore, native listeners showed greater improvement for silent-center and onset-only tones when context was given. This indicated that syllable-extrinsic information was particularly useful to the native listeners when syllable-intrinsic information was largely unavailable. Non-native listeners showed no such dramatic improvement for these two types of syllables, suggesting that extrinsic information provide by context did not make a difference between syllables rich or impoverished in intrinsic information. Taken together, the lack of impact of extrinsic information indicated that non-native listeners relied primarily on syllable-intrinsic information for tone identification.

With a few exceptions, the analyses by block showed that the correct identification patterns for tone fragments could be predicted from the amount of acoustic information presented to the listeners. Several observations are noteworthy and consistent with aforementioned interpretations of the data. In particular, for the native listeners, intact and center-only syllables were identified equally accurately, although the reaction time measure revealed that the modified syllable still incurred a processing cost. In particular, adding context to single-speaker stimuli improved identification of silent-center tones such that their accuracy was comparable to the center-only tones. Finally, multiple-speaker stimuli presented in isolation slowed down reaction to silent-center and onset-only tones such that they were no longer comparable to center-only tones.

On the non-native listeners' side, although the accuracy pattern resembles that of the native listeners, reaction time likewise revealed that the non-native listeners were particularly slow in responding to silent-center and onset-only tones. These findings showed that reaction time could be a useful measure to reveal processing differences otherwise not shown in the accuracy measure.

The identification of specific tones across listening conditions and acoustic modifications appeared to show a consistent pattern as well. For both native and non-native listeners, Tone 4 was invariably the most accurate tone and most resistant to acoustic modification. This finding replicated Lee et al. (2008) and Lee et al. (submitted), who reported that the onset of Tone 4 was acoustically most distinct from other tones. The acoustic data in the current study also showed that Tone 4 has the highest average f_0 in the first six pitch periods. For the non-native listeners, Tone 4 is also the only tone that resembles an English intonation contour (Broselow, Hurtig, & Ringen, 1987). These observations may explain the overall high identification accuracy for Tone 4.

On the other hand, Tone 2 was one of the least accurate tones in the current data. This finding also replicated Lee et al. (2008) and Lee et al. (submitted) and is consistent with the finding that intact Tone 2 was the most difficult tone to identify (Broselow et al., 1987; Wang, Spence, Jongman, & Sereno, 1999). Importantly, Lee et al. (submitted) proposed that the source of Tone 2 identification difficulty is different between native and non-native listeners. The confusion pattern analyses in Lee et al. (submitted) showed that native listeners showed a Tone 3 bias in the Tone 2-Tone 3 confusion while the non-native listeners did not show such a bias. It was proposed that native listeners were looking for positive evidence for Tone 2 (i.e., the rising f_0). When the evidence of f_0 rising was not available due to missing fragments, native listeners would treat the low onset as Tone 3. Non-native listeners, on the other hand, did not or could not use this strategy when fragments were missing. Therefore their Tone 2-Tone 3 confusion did not favor the Tone 3 response. The low accuracy of Tone 2 identification here is consistent with this interpretation.

There is another noteworthy difference between the native and non-native listeners. For the native listeners, Tones 4 and 3 always appeared in a group and so did Tones 1 and 2. For the non-native listeners, in contrast, Tones 4 and 1 formed a group and so did Tones 2 and 3. Recall from the acoustic analyses that Tones 1 and 4 begin with high f_0 onsets and Tones 2 and 3 begin with low f_0 onsets. Given the high-low f_0 onset distinction among the tones, the observation here is that the two most accurate tones for the native listeners included a high- and a low-onset tone, while the two most accurate tones for the non-native listeners were both high-onset tones. This observation implies that the native listeners were capable of making the high-low distinctions, which was not evident for the non-native listeners. As noted, the high-low tone judgment is necessarily relative and will have to take into consideration the f_0 range of a speaker. It was also speculated that context in the current task was primarily used to establish a speaker-specific reference frame. Viewed this way, the tone grouping difference between native and non-native listeners is also consistent with the finding that native listeners were more efficient at using context for speaker normalization.

5. Conclusion

The study reported here showed that native and non-native listeners differed in how they dealt with fragmented Mandarin tones produced by single vs. multiple speakers in isolation or in context. In general, the native listeners were able to make use of information in the context to facilitate tone identification from partial acoustic input. This was shown by the greater facilitation of context in the multiple-speaker set and the substantial improvement of silent-center and onset-only tone identification when context was present. In contrast, non-native listeners showed essentially the same response patterns across speaker and context conditions, indicating their focus on syllable-intrinsic information for tone identification. These results contributed to our further understanding of the nature of lexical tone processing by native and non-native listeners.

ACKNOWLEDGMENTS

We thank Ning Zhou, Ann Marie Christy, and Li Xu for assistance in data processing, and Ya-Ting Shih for assistance in administering the experiment.

REFERENCES

Broselow, E., Hurtig, R. R., and Ringen, C. (1987). The perception of second language prosody. In G. Ioup and S.H. Weinberger (Eds.), *Interlanguage Phonology: The Acquisition of a Second Language Sound System*. Cambridge, MA: Newbury House Publishers.

Connine, C. M., Clifton, C., Jr., and Cutler, A. (1987). Effects of lexical stress on phonetic categorization. *Phonetica, 44,* 133-146.

Fox, R. A., and Qi, Y.-Y. (1990). Context effects in the perception of lexical tones. *Journal of Chinese Linguistics*, *18*, 261-284.

Fox, R. A., and Unkefer, J. (1985). The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics*, *13*, 69-90.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110-125.

Gottfried, T. L., and Suiter T. L. (1997). Effects of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, *25*, 207-231.

Johnson, K. A. (2005). Speaker normalization in speech perception. In D. B. Pisoni and R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363-389). Malden, MA: Blackwell Publishing.

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98-104.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, *11*, 373-382.

Lee, C.-Y., Tao, L., and Bond, Z. S. (2008). Identification of acoustically modified Mandarin tones by native listeners. *Journal of Phonetics*. doi:10.1016/j.wocn.2008.01.002.

Lee, C.-Y., Tao, L., and Bond, Z. S. (submitted). Identification of acoustically modified Mandarin tones by non-native listeners. Submitted to *Language and Speech*.

Lin, T., and Wang, W. S.-Y. (1984). “Shengdiao ganzhi wenti.” (The issue of tone perception). *Zhongguo Yuyan Xuebao (Bull. Chinese Linguistics), 2*, 59-69

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 433-443.

Mertus, J. A. (2000). *The Brown Lab Interactive Speech System*. Brown University.

Moore, C. B., and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, *102*, 1864-1877.

Pisoni, D. B., and Remez, R. E. (2005). *The Handbook of Speech Perception*. Malden, MA: Blackwell Publishing.

Sebastián-Gallés, N. (2005). Cross-language speech perception. In D. B. Pisoni and R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 546-566). Malden, MA: Blackwell Publishing.

Shih, C. (1987). The phonetics of the Chinese tonal system. *Bell Laboratories Technical Memorandum.*

Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*, 695-705.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*, 3649-3658.

Wong, P. C. M., and Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*, 413-421.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, *95*, 2240-2253.