

Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study

Chao-Yang Lee^{a)}

School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701

(Received 28 January 2008; revised 8 November 2008; accepted 24 November 2008)

Lexical tone identification relies primarily on the processing of F0. Since F0 range differs across individuals, the interpretation of F0 usually requires reference to specific speakers. This study examined whether multispeaker Mandarin tone stimuli could be identified without cues commonly considered necessary for speaker normalization. The *sa* syllables, produced by 16 speakers of each gender, were digitally processed such that only the fricative and the first six glottal periods remained in the stimuli, neutralizing the dynamic F0 contrasts among the tones. Each stimulus was presented once, in isolation, to 40 native listeners who had no prior exposure to the speakers' voices. Chi-square analyses showed that tone identification accuracy exceeded chance as did tone classification based on F0 height. Acoustic analyses showed contrasts between the high- and low-onset tones in F0, duration, and two voice quality measures (F1 bandwidth and spectral tilt). Correlation analyses showed that F0 covaried with the voice quality measures and that tone classification based on F0 height also correlated with these acoustic measures. Since the same acoustic measures consistently distinguished the female from the male stimuli, gender detection may be implicated in F0 height estimation when no context, dynamic F0, or familiarity with speaker voices is available. © 2009 Acoustical Society of America. [DOI: 10.1121/1.3050322]

PACS number(s): 43.71.An, 43.71.Bp, 43.71.Es [AJ]

Pages: 1125–1137

I. INTRODUCTION

A. Overview

Dealing with speaker variability is an integral part of spoken language processing (Johnson, 2005). To uncover the linguistic representation intended by a speaker, a listener has to disentangle the linguistic information from speaker-specific information in the acoustic signal. The speaker variability issue is pertinent to lexical tone perception, which relies primarily on the detection of fundamental frequency (F0) (Abramson, 1978; Gandour and Harshman, 1978; Howie, 1976; Tseng, 1981). Since F0 range differs across individuals, absolute F0 values of a particular tone may vary by speaker. For example, a phonologically low tone produced by a female speaker could be acoustically equivalent to a phonologically high tone for a male speaker. What is considered high or low, then, has to be determined with respect to a speaker's F0 range.

Dealing with acoustically impoverished stimuli is also a common challenge for speech perception. Studies using degraded acoustic stimuli have revealed contextual and secondary cues to lexical tone identification (Abramson, 1972; Blicher *et al.*, 1990; Gottfried and Suiter, 1997; Lee, 2000; Lee *et al.*, 2008; Liu and Samuel, 2004; Whalen and Xu, 1992; Xu, 1994). For example, Gottfried and Suiter (1997) showed that "silent-center" Mandarin tones, where the majority of F0 information is missing, could be identified as accurately as intact tones. Overall, these studies indicate that listeners are

capable of integrating the limited F0 information in the signal or resorting to secondary cues to uncover tone identity when F0 information is not available.

How well can listeners cope with the dual challenge of speaker variability and limited acoustic input in tone perception? This article reports a perceptual experiment and acoustic analyses that examined the roles of speaker variability and limited acoustic input in Mandarin tone identification. The research question is whether tones produced by multiple speakers can be identified without prior exposure to the speakers' voices, when syllable-internal F0 information is reduced to a minimum and when the syllable-external context is not available for speaker normalization. To this end, multispeaker, naturally produced Mandarin tones were digitally processed such that only the onset consonant and the first six glottal periods remained in the stimuli. Acoustic analyses showed that the signal processing effectively removed the dynamic F0 information that distinguishes the four Mandarin tones (Tone 1: level; Tone 2: rising; Tone 3: dipping; Tone 4: falling). While these tones may still be identified based on F0 height contrasts (Tones 1 and 4: high; Tones 2 and 3: low), the height judgment depends on the knowledge of a speaker's F0 range, which should be difficult to estimate without prior exposure or an external context. If, however, listeners can identify the tones from these incomplete stimuli with accuracy exceeding chance, the question is what in the acoustic signal enables them to do so. The basis of the tone identification performance will be explored by acoustic analyses.

B. Speaker variability in lexical tone perception

Information about a speaker's F0 range can be obtained from a sentential context (Leather, 1983; Moore and Jong-

^{a)}Electronic mail: leec1@ohio.edu

man, 1997; Wong and Diehl, 2003). For contour tones presented in isolation (such as the Mandarin rising, dipping, and falling tones), speaker variability may not be a problem because these tones have distinct dynamic F0 patterns and may not require reference to F0 height for identification (Moore and Jongman, 1997). However, F0 height judgment is relevant for the perception of noncontour tones such as the Cantonese level tones (Wong and Diehl, 2003). F0 height may also be implicated in the perception of contour tones that have similar F0 patterns but are located in different registers of a speaker's F0 range. Even for tones in connected speech, F0 height estimation may still be relevant because the canonical F0 contour of a tone often changes substantially as a result of tonal coarticulation (Xu, 1994, 1997). For example, the Mandarin dipping tone becomes phonetically a low tone in nonfinal positions. Consequently, the low tone may contrast with the high tone (a level tone) only in F0 height. Under these circumstances, syllable-intrinsic F0 patterns may not provide sufficiently distinctive information and a listener will have to resort to F0 height for tone identification.

There is ample evidence that F0 height is involved in the perception of Mandarin tones even in isolation (Gandour, 1983; Gandour and Harshman, 1978; Massaro *et al.*, 1985). The role of F0 height implies the use of F0 range information. Since the tone stimuli used in these studies were presented in isolation, F0 range information could not have come from an external context. However, the synthetic tone stimuli used in these studies included contour tones; therefore the listeners could have estimated F0 height by calibrating F0 range syllable internally. The listeners could also have developed familiarity with the synthetic voice and thus the F0 range through repeated exposure to the stimuli (Honorof and Whalen, 2005; Nygaard and Pisoni, 1998; Palmeri *et al.*, 1993). Since only one speaker was modeled in these studies, it is not known whether F0 height could be judged when multiple speakers are present.

Other studies presented tones in carrier phrases of various F0 levels as a way of simulating different speakers (Fox and Qi, 1990; Leather, 1983; Lin and Wang, 1984; Moore and Jongman, 1997; Wong and Diehl, 2003). Leather (1983) and Moore and Jongman (1997) showed that synthesized Mandarin tone stimuli were identified with reference to the perceived F0 range obtained from naturally produced carrier phrases. Wong and Diehl (2003) similarly found that a naturally produced Cantonese tone stimulus was identified as different tone categories depending on the F0 height of synthesized carrier phrases. Importantly, by using level tone stimuli only, Wong and Diehl (2003) eliminated the potential contribution of syllable-intrinsic dynamic F0 patterns to the F0 range estimation. Since no syllable-internal dynamic F0 information was available, F0 height estimation must have been achieved by obtaining information from the carrier phrases.

If F0 range information can only be derived from an external context, removing the context should effectively prevent F0 range estimation, which should compromise tone identification that requires F0 height information. However, there is some indication that multispeaker, isolated tones

lacking dynamic F0 information can be identified without context. In Wong and Diehl's (2003) Experiment 1, the three Cantonese level tones (high, mid, and low) produced by seven speakers were presented in isolation. The results showed that tone identification was more accurate when the stimulus presentation was blocked by the speaker (80.3%) than when it was mixed across the speakers (48.6%). The accuracy values suggest that the listeners were able to identify the F0 level in these isolated, multispeaker tone stimuli beyond chance (33.3%). However, familiarity with voices could have contributed to the results. In particular, the listeners in the experiment heard the stimulus set 12 times. Given that the tone stimuli span across a substantial F0 range used for Cantonese tones, it is likely the multiple exposures allowed the listeners to learn the F0 range of the speakers (Honorof and Whalen, 2005; Palmeri *et al.*, 1993; Nygaard and Pisoni, 1998).

A more stringent test for speaker normalization without context, then, is to present isolated, multispeaker tone stimuli lacking dynamic F0 information without repeated exposure. No data currently exist on lexical tone materials, but Honorof and Whalen (2005) showed that English-speaking listeners were able to locate an F0 reliably within a speaker's F0 range without context or prior exposure to a speaker's voice. In their study, isolated vowel [a] tokens, produced by 20 English speakers with varying F0s, were presented to listeners to judge where each token was located in the speakers' F0 ranges. The results showed significant correlations between the perceived F0 location and the actual location in the speakers' F0 ranges. Voice quality and gender detection were implicated (Honorof and Whalen, 2005). If listeners can demonstrate such an ability for pitches that are not linguistically significant, this ability could also be implicated in the perception of lexical tones. Considering the lexically contrastive function of F0 in tone languages, the ability to locate within an F0 range isolated pitches produced by unfamiliar voices can be quite useful.

C. Tone identification from incomplete acoustic input

Since three of the four Mandarin tones are contour tones with dynamic F0 patterns, the issues of F0 height estimation and speaker normalization may not seem relevant for Mandarin tone identification. That is, Mandarin listeners should be able to rely on dynamic F0 information for tone identification. Nonetheless, research has consistently shown that detection of F0 height is implicated in Mandarin tone perception (Fox and Qi, 1990; Gandour, 1983; Gandour and Harshman, 1978; Leather, 1983; Lin and Wang, 1984; Massaro *et al.*, 1985; Moore and Jongman, 1997). To isolate the role of F0 height estimation in tone identification, a potentially useful research strategy is to use brief tone segments extracted from intact tones as stimuli such that dynamic F0 information can be minimized or removed. Although no study has been carried out that specifically examined speaker normalization in this way, there are data on how well brief tone segments can be identified. Whalen and Xu (1992) examined Mandarin tone identification from segments ranging from 40 to 100 ms extracted from various parts of a syllable.

Tone segments that do not have much F0 change were most often heard as Tone 1 (level). A low, unchanging F0 was often heard as Tone 3 (dipping). Recall that a nonfinal Tone 3 is usually realized as a low tone. These results therefore suggest that the listeners were referring to F0 height when dealing with the brief tone segments. Whalen and Xu (1992) also predicted that very short syllables such as those in running speech would show more of such “register tendency.” Sensitivity to register, as noted, implies the ability to locate a particular pitch within a speaker’s F0 range.

Gottfried and Suiter (1997) showed that isolated Mandarin tone stimuli excised from the first six glottal pulses of a syllable could be identified quite accurately (Tone 1:91%; Tone 2:28%; Tone 3:65%; Tone 4:32%). This finding was replicated by Lee *et al.* (2008) with a similar procedure but a different set of Mandarin syllables (Tone 1:68%; Tone 2:46%; Tone 3:87%; Tone 4:90%). The contrast in the actual accuracy values from these two studies could be attributed to the differences in stimuli, participants, and response format. Although it has been established that isolated Mandarin tones do not require the entire syllable to identify (Tseng, 1981; Yang, 1992), the accuracy values reported by Gottfried and Suiter (1997) and Lee *et al.* (2008) are still surprisingly high. In particular, the F0 contours remaining in the six glottal pulses were basically flat and did not show distinct patterns among the four tones (Lee *et al.*, 2008). Even if there were distinct F0 contours acoustically, the dynamic character of the pitch contour cannot be perceived at such a short duration (Greenberg and Zee, 1979). On the other hand, the acoustic analyses of Lee *et al.* (2008) did show distinct groupings of the four tones in F0 height: the high-onset tones included Tones 1 and 4; the low-onset tones included Tones 2 and 3. These acoustic data are consistent with the finding that tones belonging to the same height group were more confusable perceptually. It appears that the listeners were indeed sensitive to F0 height contrasts when dynamic F0 information was not available. Again, being able to judge F0 height implies the ability to locate pitches in a speaker’s F0 range.

The sensitivity to F0 height, however, was not observed in Wu and Shu’s (2003) data. Using the gating paradigm (Grosjean, 1996), Wu and Shu (2003) presented isolated Mandarin syllable fragments in 40 ms increments, starting with an 80 ms gate. The task for the participants was to propose a monosyllabic word candidate at each gate that best matched the acoustic input. Analyses of the candidates proposed at each gate showed that the most common tone error for Tone 1 stimuli was Tone 4 up to 120 ms of input, which is consistent with the confusion patterns from the previous three studies (Whalen and Xu, 1992; Gottfried and Suiter, 1997; Lee *et al.*, 2008). For stimuli generated from the other three tones, however, the most common error was invariably Tone 1. While these results are consistent with Greenberg and Zee’s (1979) finding that tone contours cannot be perceived with durations shorter than 130 ms, the predominant Tone 1 response, even for the low-onset tones (Tones 2 and 3), did not indicate any sensitivity to F0 height.

As with earlier studies, only one speaker was used to generate the stimuli in Whalen and Xu (1992), Gottfried and Suiter (1997), Lee *et al.* (2008), and Wu and Shu (2003). For

the three studies that did show signs of sensitivity to F0 height, the listeners had ample opportunity to become familiar with the speaker’s voice through repeated exposures. Familiarity, as noted, could contribute to F0 range estimation. Therefore it is not clear if the high accuracy reported in some of the studies truly reflects the ability to detect F0 height. Nonetheless, these studies showed that the use of brief, gated tone stimuli can be an effective way of neutralizing dynamic F0 information in Mandarin tones.

D. Summary and predictions

The available evidence suggests that lexical tone perception involves processing F0 information with reference to a speaker’s F0 range (Leather, 1983; Moore and Jongman, 1997; Wong and Diehl, 2003). However, how speaker normalization is accomplished remains underspecified. The acoustic basis for the perceptual ability also remains unsubstantiated. The finding that nonlinguistic, isolated pitches produced by an unfamiliar voice can be reliably located within a speaker’s F0 range (Honorof and Whalen, 2005) invites the question of whether a similar ability is implicated in the perception of lexically contrastive tones.

The present study aims to address this issue by examining the identification of brief portions of Mandarin tones and the acoustic characteristics of the tone stimuli. Only six glottal pulses were preserved in the tone stimuli such that syllable-internal dynamic F0 information was neutralized. The stimuli were presented in isolation such that no contextual information was available. Each of the multispeaker tone stimuli was presented only once such that familiarity with speaker voice was minimized. Since dynamic F0, context, and familiarity are known to contribute to speaker normalization, eliminating all of them from the stimuli should make speaker normalization very difficult. Consequently, tone identification from these multispeaker stimuli should not exceed chance level. If so, this result will indicate that dynamic F0, context, and/or familiarity are indeed necessary for speaker normalization.

If, however, these tone stimuli can be identified with accuracy exceeding chance, listeners must be able to obtain useful information from the isolated stimuli for speaker normalization. One possibility is that some acoustic properties in the stimuli covary with F0 to facilitate F0 height estimation (Honorof and Whalen, 2005). Previous research has shown that secondary cues such as duration and amplitude contour can be used for lexical tone identification when F0 information is compromised (Abramson, 1972; Blicher *et al.*, 1990; Liu and Samuel, 2004; Whalen and Xu, 1992). Voice quality has also been proposed as the basis for the ability to locate a pitch in a speaker’s F0 range (Honorof and Whalen, 2005; Swerts and Veldhuis, 2001). In particular, voice quality may covary with F0, or it may be used for gender detection as a first pass for F0 height estimation. It is known that listeners are able to detect speaker gender from units of speech ranging from the sentence to individual sounds (Childers and Wu, 1991; Ingemann, 1968; Lass *et al.*, 1978). Honorof and Whalen (2005) also reported positive correlations between perceived pitch locations and gender-related

estimates, indicating the potential role of gender detection in F0 height estimation. Considering the female-male difference in average F0 (Peterson and Barney, 1952) and voice quality (Hanson, 1997; Hanson and Chuang, 1999), successful gender identification could mediate F0 height judgments. The potential contribution of these measures will be explored in the acoustic analyses.

Another possibility is that F0 height can be estimated based on some internal templates of speaking F0 range, which are formed based on experience with the prevailing F0 range of the speakers in a linguistic community (Dolson, 1994; Honorof and Whalen, 2005). Dolson (1994) noted that despite anatomical variabilities across individuals, the speaking F0 range is not wildly variable within a linguistic community. His survey showed that 80% of the male speakers within a linguistic community have an average speaking F0 within three semitones of the group mean, and the female speakers' F0 range is even narrower. The idea is that listeners can acquire an internal representation of pitch classes based on the prevailing speaking F0 range in their linguistic community, and that pitch perception and production are both influenced by this representation (Deutsch, 1991; Deutsch, *et al.*, 1999, 2004; Deutsch *et al.*, 1990). If the speaking F0 range in a linguistic community is sufficiently constrained, listeners may be able to locate a particular F0 based on these stored pitch templates.

II. TONE IDENTIFICATION EXPERIMENT

A. Method

1. Materials

The Mandarin syllable *sa*, produced with all four tones by 16 female and 16 male Beijing Mandarin speakers (age range: 23–35 years), was selected to generate the stimuli for the experiment. The syllables were drawn from an existing database collected at Ohio University in 2004 that includes the Mandarin syllables *sa*, *xia*, and *sha*. The *sa* syllable was chosen because the same materials were to be used in a subsequent study that would investigate brief tone identification by English-speaking listeners. To that end, the intention was to select a syllable that exists in both languages to avoid potential interference from segmental structure. Among the three fricatives, the alveolar fricative is phonologically closest to an actual phoneme in English. In contrast, the articulatory and acoustic characteristics of the other two Mandarin fricatives differ more substantially from the English palatoalveolar fricative (Ladefoged, 1996; Stevens *et al.*, 2004).

The recordings were made in a sound-treated booth with an Audio-technica AT825 field recording microphone connected through a preamplifier and A/D converter (USBPre microphone interface) to a Windows personal computer. The speakers were instructed to read the syllables in citation form. The recordings were digitized with the Brown Lab Interactive Speech System (BLISS, Mertus, 2000) at 44.1 kHz with 16 bit quantization. Each syllable was identified from the BLISS waveform display, excised from the master file, and saved as an individual audio file. The peak amplitude was normalized across syllables.

Each *sa* syllable was digitally processed with BLISS such that the stimuli included only the fricative consonant and the first six glottal periods. The cut was always made at a zero crossing. There were no perceptible clicks as a result of the signal processing; therefore no further tapering procedure was applied. A total of 128 stimuli (4 tones \times 32 speakers) were used in the experiment.

2. Participants

Forty Beijing Mandarin speakers were recruited from the Ohio University community to participate in the experiment with cash compensation. The participants included 27 females (mean age=24 years, SD=4.5) and 13 males (mean age=21 years, SD=2.6). All spoke Mandarin on a daily basis and none reported any speech or hearing difficulties. Twenty-one participants reported speaking some dialect of Chinese other than Mandarin, but all identified Mandarin as their native language.

3. Procedure

The stimuli were imported to AVRrunner, the subject-testing program in BLISS, for stimulus presentation and response data acquisition. To minimize the impact of familiarity with individual speaker voices, the 128 stimuli produced by the 32 speakers were assigned to four blocks such that each block included only one stimulus from a speaker. That is, each block had 32 stimuli and all stimuli were produced by different speakers. Within each block, the number of male and female speakers was balanced (i.e., 16 females and 16 males) as was the number of the four tones (i.e., eight stimuli for each of the four tones). No two participants received the same order of presentation. The order of presentation for the blocks was also randomized. The interstimulus interval was 5 s. Eight practice stimuli were given prior to the experiment to familiarize the participants with the procedure and response format. The practice stimuli, also excised *sa* syllables, were recorded by a female and a male speaker not used in the actual experiment.

Participants were tested individually in a sound-treated room. They listened to the stimuli through a pair of KOSS R80 headphones connected to a personal computer. The participants were told they would be listening to the syllable *sa* with all four tones produced by 32 female and male speakers. They were also told the syllables had been digitally processed such that only the very beginning of the syllables was audible. Their task was to identify the tone of each stimulus by pressing the buttons labeled "1," "2," "3," and "4" on the computer keyboard, representing the four Mandarin tones. All participants indicated they were familiar with the convention of designating Mandarin tones with the numbers. The participants were further told that they had 5 s to respond to each stimulus and that their response would be timed.

B. Results

To evaluate the null hypothesis that tone identification responses were not related to the stimulus tones, the expected and actual responses were tabulated to generate con-

TABLE I. Contingency tables showing the frequency counts of observed and expected (in parentheses) tone identification responses.

Stimulus	Female stimuli Response			
	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	281 (234.75)	102 (109.75)	59 (120)	194 (171.5)
Tone 2	239 (235.119)	133 (109.923)	140 (120.189)	125 (171.77)
Tone 3	187 (235.119)	108 (109.923)	213 (120.189)	129 (171.77)
Tone 4	232 (234.012)	96 (109.405)	68 (119.623)	238 (170.961)

Stimulus	Male stimuli Response			
	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	268 (226.821)	116 (106.416)	107 (150.132)	144 (151.631)
Tone 2	228 (227.536)	103 (106.751)	180 (150.605)	126 (152.108)
Tone 3	154 (226.464)	88 (106.249)	228 (149.895)	164 (151.392)
Tone 4	258 (227.179)	119 (106.584)	86 (150.368)	173 (151.869)

tingency tables for χ^2 tests. Table I shows the numbers of expected and actual tone responses to the tone stimuli. The expected frequencies were calculated as follows: Suppose E_{ij} is the expected frequency for the cell in row i and column j , R_i and C_j are the corresponding row and column totals (marginal totals), and N is the total number of observations, then $E_{ij}=R_iC_j/N$ (Howell, 1999; Cohen, 1996).

Separate χ^2 tests were conducted for female and male stimuli to test the null hypothesis that the responses were independent of the stimuli. Both tests showed that the null hypothesis should be rejected: female, $\chi^2(9, N=2544)=206.96, p<0.0001$; male, $\chi^2(9, N=2542)=135.66, p<0.0001$. (The unequal N resulted from 16 and 18 missing responses in the female and male data, respectively.) These results indicate that the listeners' tone identification responses were not random or totally unrelated to the tone stimuli. That is, the overall accuracy was different from chance level performance.

A second set of contingency tables was generated based on specific tones. To evaluate Tone 1 identification, for example, both the stimuli and responses were coded as Tone 1 or non-Tone 1. Therefore, all responses could be classified into "hit" (Tone 1 stimulus \rightarrow Tone 1 response), "miss" (Tone 1 stimulus \rightarrow non-Tone 1 response), "false alarm" (non-Tone 1 stimulus \rightarrow Tone 1 response), and "correct rejection" (non-Tone 1 stimulus \rightarrow non-Tone 1 response). For the remaining three tones, the same coding procedure was applied to generate similar 2×2 contingency tables. As before, χ^2 tests were performed to test the null hypothesis that the responses were not related to the stimuli. For the female stimuli, all four χ^2 tests indicate that the null hypothesis should be rejected: Tone 1, $\chi^2(1, N=2544)=19.26, p<0.0001$; Tone 2, $\chi^2(1, N=2544)=7.81, p<0.01$; Tone 3, $\chi^2(1, N=2544)$

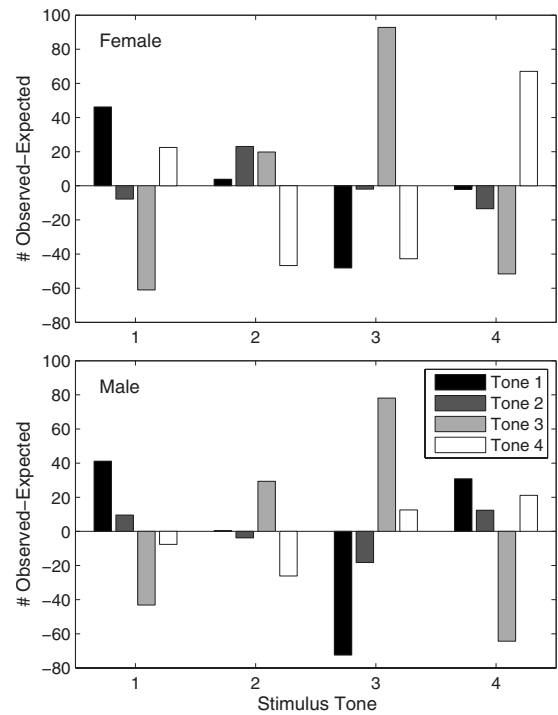


FIG. 1. Difference between the numbers of observed and expected tone responses as a function of stimulus tone in the tone identification experiment.

$=117.85, p<0.0001$; Tone 4, $\chi^2(1, N=2544)=47.94, p<0.0001$. For the male stimuli, all but the Tone 2 result indicate that the null hypothesis should be rejected: Tone 1, $\chi^2(1, N=2542)=15.5, p<0.0001$; Tone 2, $\chi^2(1, N=2542)=0.21, n.s.$; Tone 3, $\chi^2(1, N=2542)=71.01, p<0.0001$; Tone 4, $\chi^2(1, N=2542)=5.15, p<0.05$. Overall, these results show that tone identification performance was above chance level except for the responses to Tone 2 stimuli produced by the male speakers.

A final set of χ^2 tests was conducted based on F0 height of the tones: high-onset tones include Tones 1 and 4; low-onset tones include Tones 2 and 3. As the acoustic analyses will show, dynamic F0 information was absent from these brief stimuli, but the F0 height difference remained between the high- and low-onset tones. If the identification data show that the high-low distinction can be detected beyond chance, the ability to estimate F0 height is likely the basis for the tone identification performance. Indeed, when both the stimuli and responses were coded by F0 height, χ^2 tests showed that the high-low judgments were not random: for the female stimuli, $\chi^2(1, N=2544)=121.95, p<0.0001$; for the male stimuli, $\chi^2(1, N=2542)=47.77, p<0.0001$. These results indicate that the listeners were capable of judging the high-low distinction above chance.

The contingency tables also provide information about the confusion patterns among the four tones. To facilitate the interpretation of the patterns, Fig. 1 shows the difference between the expected and actual response counts as a function of stimulus tone. Each bar represents the result of subtracting the number of expected responses from the number of actual responses in each cell of Table I. Positive or greater values indicate that the response tone is more confusable

with the stimulus tone. By contrast, the smaller or more negative a value is, the less confusable the response tone is with the stimulus tone. There are some differences between the female and male stimuli, but some common patterns can still be seen. In particular, Tone 1 was least often misidentified as Tone 3; Tone 2 was most confused with Tone 3 and least identified as Tone 4; Tone 3 was rarely misidentified as Tone 1; Tone 4 was least misidentified as Tone 3. There are also some gender-specific patterns. Specifically, while the female data indicate that Tone 1 was most often misidentified as Tone 4, the male data show that Tone 4 was most often misidentified as Tone 1. These results suggest there might be gender differences in the stimuli, which will be examined in the acoustic analyses. Nonetheless, all these results are consistent with the observation that tone stimuli sharing a similar F0 height are more confusable, indicating that the high-low distinction can be detected by the listeners.

C. Discussion

When presented with multispeaker, isolated tone stimuli with six glottal periods, Mandarin listeners were able to identify the tones of the stimuli with accuracy exceeding chance. With the exception of Tone 2 produced by the male speakers, tone identification responses were not random but were contingent on the tone stimuli, indicating the ability to distinguish the target tone from the nontarget tones. With only six glottal periods in a stimulus, it is unlikely that dynamic F0 information was available for tone identification. Rather, detection of F0 height is most likely the basis for tone identification. This interpretation is further evaluated by acoustic analyses reported in Sec. III.

These results suggest that F0 height detection is the basis for the tone identification performance. Given that the tone stimuli were produced by 32 speakers and that the F0 height of a tone is likely to vary across individuals, it can be inferred that some sort of speaker normalization was involved in the tone identification process. Since no context was given, whatever allowed the speaker normalization must have come from the stimuli internally. Familiarity with individual speaker voices is unlikely to have contributed to the normalization since each tone stimulus was presented only once. Taken together, it is reasonable to conclude that some acoustic characteristics in the stimuli facilitated the speaker normalization and thus the tone identification.

As with most phonetic distinctions, there are multiple acoustic cues to tonal contrasts (e.g., Whalen and Xu, 1992). Since dynamic F0 information was presumably compromised in the stimuli, it is conceivable that some of the secondary cues such as duration and amplitude might be implicated. Acoustic analyses were conducted on the tone stimuli to explore the contribution of these cues. The acoustic analyses were also intended to verify several assumptions made earlier such as the lack of dynamic F0 in the stimuli, the high-versus low-onset distinction among the four tones, and the existence of speaker variability in F0 height. Finally, the acoustic analyses could also explicate the gender differences observed in the perceptual results. In particular, the χ^2 test for male Tone 2 stimuli failed to reach significance; the over-

all accuracy for male stimuli was also lower. These results suggest the tonal distinctions might not be as clear in the male stimuli. These observations will be evaluated by the acoustic analyses.

The Tone 2 stimuli in the current study appear to have a disadvantage compared to the other tones. In particular, the contingency table analyses showed the Tone 2 stimuli produced by the male speakers were the only stimuli among the four tones that failed to be identified with accuracy exceeding chance. A potentially confounding factor is lexical status/frequency. In particular, the syllable *sa* with Tone 2 happens to be an accidental gap in the Mandarin lexicon; that is, this syllable-tone combination is not associated with a real word in the language. Even when all the syllables that begin with the *sa* sequence are considered (*sa*, *sai*, *san*, *sang*, and *sao*), none of the syllable-tone combinations is associated with a real word (Wang, 1986). Given the demonstrated influence of lexical status on Mandarin tone categorization (Fox and Unkefer, 1985), the Tone 2 disadvantage could have resulted from the accidental gap. Alternatively, since Gottfried and Suiter (1997) and Lee *et al.* (2008) also reported the lowest accuracy and/or longest reaction time for onset-only Tone 2, it is equally plausible that the low accuracy for Tone 2 reflects the acoustic characteristics of the stimuli or global distributional patterns of the four tones in the language. Nonetheless, a syllable that is balanced in lexical frequency across all four tones should be used in future studies.

III. ACOUSTIC ANALYSES

Acoustic analyses were conducted on the F0, duration, amplitude, and voice quality of the stimuli. Examining F0 is an obvious choice since it is the primary acoustic correlate of Mandarin tones (e.g., Abramson, 1978). Duration is a direct measure of the amount of acoustic input and has been shown to be a cue to tone identification when F0 is not available (e.g., Liu and Samuel, 2004; Whalen and Xu, 1992). There is also evidence that amplitude contours could be used to identify Mandarin tones in the absence of F0 (Whalen and Xu, 1992). Although the peak amplitude had been normalized for all the syllables prior to the silencing procedure, it is possible that some differences among the four tones could still exist in the gated stimuli.

The use of voice quality to distinguish lexical tones has been noted in several acoustic studies (Andruski, 2006; Andruski and Ratliff, 2000; Huffman, 1987). Specifically, Andruski (2006) and Andruski and Ratliff (2000) showed that the amplitude difference between the first and second harmonic (H1-H2), a measure of glottal open quotient, is distinct among the modal, creaky, and breathy tones that occupy a crowded F0 space in Green Mong. For these tones that are minimally distinct in F0, voice quality carries part of the functional load for tonal distinctions. Although voice quality is not used phonemically for Mandarin tones, isolated Tone 3 has been noted to involve glottalization for some speakers (e.g., Liu and Samuel, 2004). In addition, acoustic measures of glottal configuration have been shown to distinguish between female and male voices (Hanson, 1997; Hanson and Chuang, 1999), which could be implicated in F0 height esti-

mation (Honorof and Whalen, 2005). To explore the potential contribution of glottal configuration, acoustic measures that have been shown to reflect the status of the glottis (open quotient, F1 bandwidth, and spectral tilt) were examined.

A. Method

The F0 of the first six glottal periods was measured cycle by cycle from the BLISS waveform display. The duration of the six glottal periods was also measured from the waveform display. Root mean square (rms) amplitude was calculated using MATLAB. Specifically, the sound files were imported into MATLAB as digital samples with amplitude values ranging from -1 to 1 . The rms values were calculated for the frication noise and the six glottal periods separately. A normalized amplitude was also calculated by subtracting the amplitude of the frication noise from that of the six glottal periods.

Finally, three measures of voice quality were taken that have been shown to indicate glottal configuration. They included open quotient, F1 bandwidth, and spectral tilt (Hanson, 1997; Hanson and Chuang, 1999; Stevens, 1998). In particular, the amplitude difference between the first and second harmonic (H1-H2) is a measure of open quotient or the percent of a glottal cycle in which the glottis is open. A larger H1-H2 value indicates that the glottis is open for a significant portion of a cycle. The amplitude difference between the first harmonic and the strongest harmonic in the F1 range (H1-A1) is a measure of F1 bandwidth. A wider F1 bandwidth indicates greater acoustic loss at the glottis due to the incomplete glottal closure. The amplitude difference between the first harmonic and the strongest harmonic in the F3 range (H1-A3) is a measure of spectral tilt. A larger spectral tilt also indicates greater acoustic loss at the glottis, indicating that the glottis is never closed during a cycle. To obtain these voice quality measures, a Hamming window of 25.6 ms was placed at the beginning of acoustic periodicity for each stimulus. A power spectrum was generated with BLISS by applying discrete Fourier transform to the signal. The first harmonic (H1), second harmonic (H2), the strongest harmonic in the F1 range (A1), and the strongest harmonic in the F3 range (A3) were identified from the spectrum. The amplitude values were measured from the spectrum to derive the three measures of glottal configuration.

To evaluate whether these acoustic properties were different among the tones and between the genders, analyses of variance (ANOVAs) were conducted on the acoustic measures with stimulus tone (1, 2, 3, and 4) as a within-subject factor, gender (female and male) as a between-subject factor, and speakers as a random factor. When a main effect from an ANOVA was significant, the Bonferroni *post hoc* test was used for pairwise means comparisons to keep the familywise type I error rate at 5%.

B. Results

1. Fundamental frequency

Figure 2 shows the average F0 values for each of the six glottal periods. Each data point represents an average of 16 female or 16 male speakers. As the figure shows, all four F0

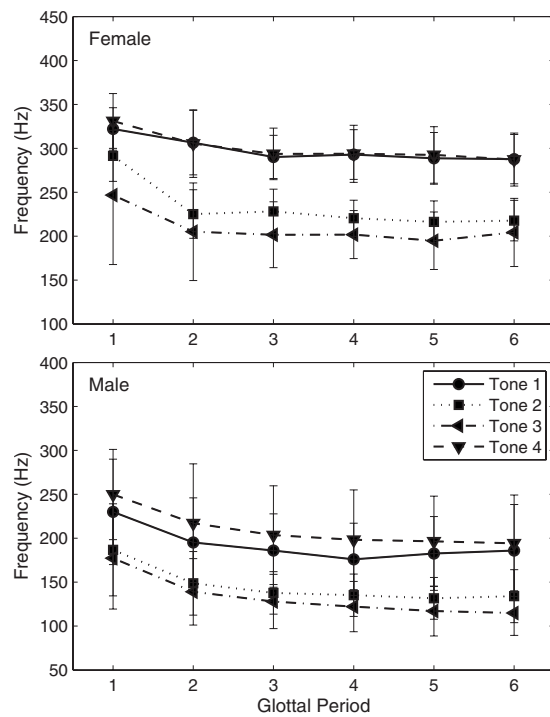


FIG. 2. Average F0 (± 1 SD) of each of the six glottal periods in the tone stimuli. Each data point represents an average of 16 female or 16 male speakers.

contours are fairly similar to each other, showing minimal contrasts. Specifically, the F0 drops somewhat during glottal cycles 1–3 and stays relatively flat during glottal cycles 3–6. As predicted, the dynamic F0 difference among the four tones has been neutralized.

For both the female and male stimuli, the four tones form two distinct groups. In particular, Tones 1 and 4 have higher F0 values than Tones 2 and 3. Despite the individual variability shown by the error bars, the high-onset tones are fairly separated from the low-onset tones. This is particularly obvious for the female speakers, where the high-onset tones show little overlap with the low-onset tones. The grouping of the four tones is consistent with the traditional description of isolated Mandarin tones (that Tones 1 and 4 start higher than Tones 2 and 3) and with the confusion patterns found in the identification experiment.

To obtain a quantitative measure of the tone and gender differences, a mean F0 value for each tone was calculated by averaging across the six glottal periods. Since the F0 from one period to the next are similar, an average across the six glottal pulses is a reasonable representation of the tone. The ANOVA on the mean F0 showed a main effect of gender, $F(1,30)=67.15$, $p<0.0001$, and a main effect of tone, $F(3,90)=119.58$, $p<0.0001$. The gender effect is expected, with female speakers showing a higher average F0 than male speakers. For the tone effect, pairwise means comparisons showed all pairwise comparisons were significant except for the Tone 1–Tone 4 comparison. Specifically, Tones 1 and 4 have a higher F0 than Tone 2, which in turn has a higher F0 than Tone 3. That is, all tone pairs, except for Tones 1 and 4, are statistically distinct from each other. The lack of difference between Tones 1 and 4 is consistent with the confusion

pattern reported earlier. Furthermore, the result that Tone 3 occupies the very low end of the F0 range and contrasts maximally with Tones 1 and 4 is also compatible with the finding that Tone 3 was identified quite accurately (Recall in the contingency table analyses, Tone 3 had the greatest positive difference between observed and expected responses for both the female and male stimuli). As Whalen and Xu (1992) also reported, a low, unchanging F0 is often heard as Tone 3. The converging evidence suggests that the listeners were able to detect a low F0 fairly well and that they were likely to call it a Tone 3.

Finally, the existence of speaker variability in F0 height is evident in Fig. 2. In addition to the variability shown by the error bars, the comparison between the female and male stimuli is particularly telling. In particular, the low-onset tones for the female speakers overlap substantially with the high-onset tones for the male speakers. Acoustically, the female “low” tones are almost identical to the male “high” tones. Nonetheless, the contingency table analyses showed that all four tones involved (female: Tones 2 and 3; male: Tones 1 and 4) were identified with accuracy beyond chance. This observation suggests the possibility that the listeners were able to detect speaker gender first and then locate the stimulus F0 within the F0 range based on the gender judgment.

2. Duration

The average duration of the six glottal periods ranges from 20 to 49 ms. For the female stimuli, Tone 1 (21 ms, SD=2), Tone 2 (26 ms, SD=2), Tone 3 (31 ms, SD=8), and Tone 4 (20 ms, SD=2). For the male stimuli, Tone 1 (33 ms, SD=8), Tone 2 (44 ms, SD=8), Tone 3 (49 ms, SD=13), and Tone 4 (31 ms, SD=8). With such a short duration, the question arises whether any durational differences can be perceived. Although no data are available on the duration difference limen (ΔT) for lexical tones, several studies using noise and pure tone burst stimuli provided information on ΔT as a function of stimulus duration (Abel, 1972; Sinnott *et al.* 1987; Dooley and Moore, 1988). The results from these studies are generally consistent (Gelfand, 1988; Yost, 2007). Inspection of Fig. 1 of Abel (1972), for example, shows that for stimuli of 10–50 ms long, ΔT is in the range of 3–5 ms. Since the duration differences found in the current study range from 0 to 19 ms, some of the differences do exceed the ΔT reported in Abel (1972). Certainly the noise/pure tone discrimination results may not generalize to lexical tones, but if the duration difference can indeed be perceived, duration could play a role in tone or gender identification.

Statistically, the ANOVA on duration showed a main effect of gender, $F(1, 30)=50.12$, $p<0.001$; a main effect of tone, $F(3, 90)=56.88$, $p<0.001$; and a significant gender \times tone interaction, $F(3, 90)=4.26$, $p<0.01$. The gender effect is expected. In particular, since the male speakers on average have a lower F0 and the period is the inverse of F0, the male duration would be longer compared to the female duration. The longer duration and presumably the greater amount of acoustic information in the male stimuli, however, did not translate to higher response accuracy. On the contrary, the identification results showed the opposite pattern,

with the female stimuli generating more accurate responses. Assuming the duration difference can be perceived, this indicates that longer duration does not necessarily provide higher information value. The contrast between the female and male stimuli, if it could be perceived, also suggests that duration could be useful for gender detection: shorter duration for females and longer duration for males. Even though the short-long contrast is also a relative one, exposure to the 128 stimuli could have provided the “context” for the judgment.

For the tone effect, all pairwise comparisons were significant except for the Tone 1–Tone 4 comparison. The result that the duration of Tones 2 and 3 is longer than that of Tones 1 and 4 can be readily predicted from their F0 contrasts. Assuming the duration difference can be perceived and using duration as an index of the amount of information, we would predict Tones 2 and 3 to be identified better than Tones 1 and 4. This prediction is only partially consistent with the identification results. In particular, Tone 3 was identified quite accurately, but Tone 2 was identified with the lowest accuracy. Finally, although the gender \times tone interaction is statistically significant, inspection of the interaction plot did not reveal any notable patterns other than those already observed in the main effects.

Finally, as noted in Sec. I, Greenberg and Zee (1979) showed that the dynamic character of the pitch contour of a tone cannot be perceived with a duration shorter than 130 ms. Since no stimulus in the current study exceeded 49 ms in duration, it is unlikely the dynamic F0 contrasts among the four tones were perceptible even if there were F0 contour differences. This is additional evidence that the signal processing effectively neutralized the F0 contour difference among the four tones.

3. Amplitude

The average rms amplitude of the six glottal periods was obtained and the normalized rms amplitude adjusted by subtracting the friction noise amplitude was calculated. As noted, amplitude was calculated on values ranging from -1 to 1 . Since the two measures essentially showed the same pattern, Fig. 3 shows only the rms amplitude. None of the ANOVAs on both measures showed an effect of tone or gender, indicating that no amplitude difference exists among the four tones or between the female and male stimuli. As noted, prior to the silencing procedure, peak amplitude had been equated across all intact syllables for the purpose of normalizing across speakers. Consequently, potential amplitude differences among the four tones, if any, may have been neutralized. Although Whalen and Xu (1992) showed that amplitude contour could be a cue for Mandarin tone identification, their results were based on signal-correlated noise stimuli of a syllable length when no F0 information was available. Whether amplitude could be of use of identification of brief stimuli with F0 information awaits further research.

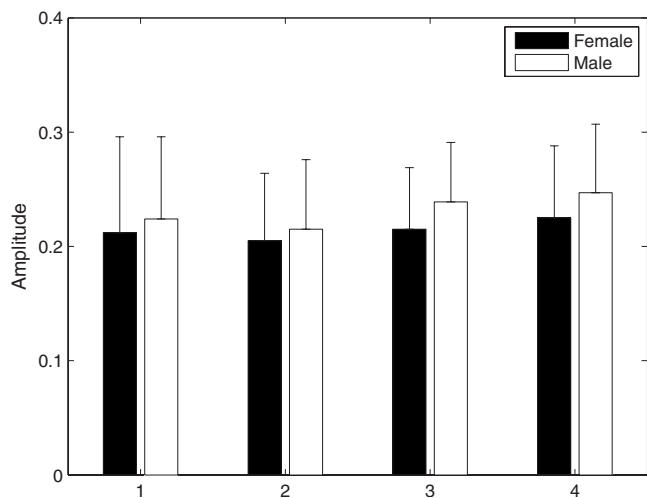


FIG. 3. Average rms amplitude (+1SD) of the six glottal periods in the tone stimuli.

4. Voice quality

Figure 4 shows the results of the three voice quality measures. For H1-H2, a measure of glottal open quotient, the ANOVA revealed no main effects but a significant gender \times tone interaction, $F(3, 90)=3.89$, $p<0.05$. Overall, H1-H2 values for the tones were more distinct in the female stimuli. Specifically, the female data showed that Tones 1 and 4 have higher H1-H2 values than Tones 2 and 3. That is, the proportion of closure during a glottal cycle is higher for Tones 2 and 3 (the low-onset tones) and lower for Tones 1 and 4 (the high-onset tones).

For H1-A1, a measure of F1 bandwidth, the ANOVA showed significant main effects of gender, $F(1, 30)=5.88$,

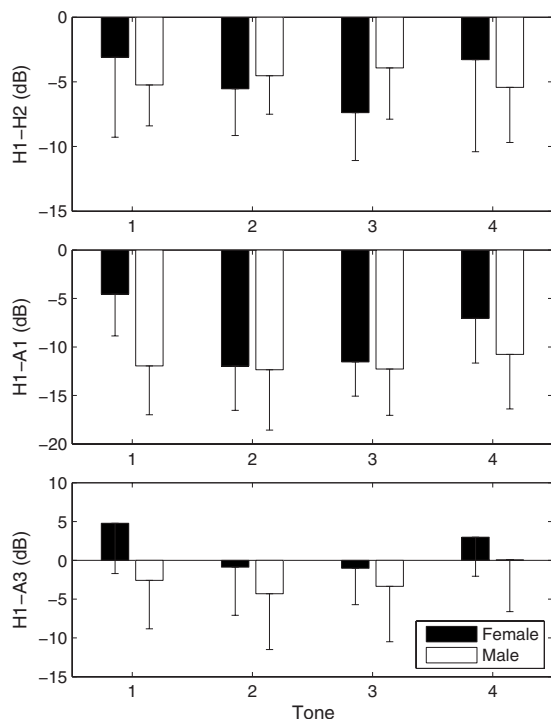


FIG. 4. Voice quality measures (-1SD) in the tone stimuli including open quotient (H1-H2), first formant bandwidth (H1-A1), and spectral tilt (H1-A3).

$p<0.05$, and tone, $F(3, 90)=8.63$, $p<0.0001$. There is also a significant tone \times gender interaction, $F(3, 90)=5.64$, $p<0.005$. Overall, the female stimuli have higher H1-A1 values than the male stimuli, indicating wider F1 bandwidth and thus greater acoustic loss at the glottis. For the tone effect, all pairwise comparisons were significant except for the Tone 1-Tone 4 comparison and the Tone 2-Tone 3 comparison. That is, the F1 bandwidth is greater for Tones 1 and 4 than for Tones 2 and 3, indicating greater acoustic loss at the glottis for Tones 1 and 4. The interaction shows that H1-A1 values were more distinct among the tones in the female stimuli than in the male stimuli.

For H1-A3, an index of spectral tilt, the ANOVA revealed significant main effects of gender, $F(1, 30)=5.87$, $p<0.05$, and tone, $F(3, 90)=6.37$, $p<0.001$. Overall, the female stimuli have higher H1-A3 values than the male stimuli, indicating larger spectral tilt and thus greater acoustic loss at the glottis for the female speech. For the tone effect, all pairwise comparisons were significant except for the Tone 1-Tone 4 comparison and the Tone 2-Tone 3 comparison. That is, spectral tilt is greater for Tones 1 and 4 than for Tones 2 and 3, again indicating greater acoustic loss at the glottis for the high-onset tones.

C. Relating acoustic and perceptual measures

The acoustic analyses showed that F0, duration, and two voice quality measures (F1 bandwidth and spectral tilt) were distinct between the high- and low-onset tones and between the female and male stimuli. These acoustic differences could be the basis for the findings from the identification experiment that listeners could identify the high- and low-onset tones above chance. In addition, the tone \times gender interaction found in duration, open quotient, and F1 bandwidth showed that the tonal contrasts were more distinct in the female stimuli. This result is also consistent with the finding that tone identification from the female stimuli was more accurate.

To evaluate the acoustic-perceptual relationship quantitatively, Pearson's correlation coefficients were derived between the seven acoustic measures (F0, duration, amplitude, normalized amplitude, open quotient, F1 bandwidth, and spectral tilt) and the two perceptual measures (accuracy of tone identification and accuracy of high-low F0 identification). Since the acoustic measures were obtained for each stimulus item, the accuracy of tone/F0 height identification was defined as the number of listeners who correctly identified the tone/F0 height of a stimulus. That is, the correlation analyses were conducted by items. Among the acoustic measures, a significant correlation was expected between F0 and duration because frequency and period are inversely related. Significant correlations were also expected among the three voice quality measures because they all involve estimation of the degree of acoustic loss at the glottis.

Table II shows a correlation matrix with all the acoustic and perceptual measures. Fisher's r to z transformation was carried out on the correlation coefficients to determine if the coefficients are significantly different from zero. The expected correlations between F0 and duration and among the

TABLE II. Correlations between acoustic and perceptual measures of tone identification. Statistically significant correlations are indicated by an asterisk (*). A total of 128 observations were used in the computation.

	Accuracy by F0 height	F0	Duration	Amplitude	Normalized amplitude	Open quotient	F1 bandwidth	Spectral tilt
Accuracy by tone	0.649*	0.225*	-0.085	-0.004	-0.015	-0.003	0.168	0.121
Accuracy by F0 height		0.431*	-0.306*	0.045	0.055	0.049	0.321*	0.216*
F0			-0.934*	-0.031	-0.014	0.083	0.451*	0.391*
Duration				0.055	0.032	-0.024	-0.396*	-0.358*
Amplitude					0.969*	-0.07	-0.079	0.038
Normalized amplitude						-0.116	-0.151	-0.005
Open quotient							0.492*	0.356*
F1 bandwidth								0.646*

three voice quality measures did turn out significant. In addition, F0 is significantly correlated with F1 bandwidth and spectral tilt. Duration is likewise correlated with the same voice quality measures. Taken together, duration, F1 bandwidth, and spectral tilt covary with F0, suggesting the listeners could have taken advantage of this covariation for F0 height estimation. Indeed, the accuracy of F0 height identification is significantly correlated with exactly these four acoustic measures.

Regression analyses were conducted to evaluate how well the acoustic measures could predict tone identification accuracy. To ensure an adequate sample size (Tabachnick and Fidell, 2001), the analyses were conducted across all four tones such that all 128 observations could be used. Since the ANOVAs on the acoustic measures reported earlier showed that the two amplitude measures did not distinguish the four tones, they are not likely to contribute to the tonal distinctions perceptually. Therefore, the two amplitude measures were not included in the regression models. Furthermore, since the correlation analyses noted earlier showed high correlation between F0 and duration, using both F0 and duration as predictors would introduce multicollinearity, which usually results in low tolerances and misleading beta weights (e.g., Cohen, 1996). It is more reasonable to enter F0 instead of duration based on the assumption that the duration difference may not be perceptible.

Two regression models were built: one with tone identification accuracy as the dependent variable and the other with tone classification based on F0 height as the dependent variable (high-onset tones include Tones 1 and 4; low-onset tones include Tones 2 and 3). In both models, the predictors included F0 and the three voice quality measures (open quotient, F1 bandwidth, and spectral tilt). In the first model, the four predictors accounted only for 6.1% of the variance and none of the regression coefficients turned out significant. In the second model, the predictors accounted for 21.1% of the variance. Specifically, F0 was the best predictor ($\beta=0.351$, $p<0.0005$). H1-A1 was the second best predictor even though the coefficient was not significant ($\beta=0.226$, $p=0.06$). These results indicate that F0 is the only statistically significant predictor of tone classification based on F0 height, suggesting that the listeners were able to use F0 height information for the classification. This observation is

also consistent with the hypothesis that listeners were using the covariation among F0 and two of the voice quality measures (F1 bandwidth and spectral tilt) found in the correlation analyses for F0 height estimation.

IV. GENERAL DISCUSSION

The major finding from the present study is that tone identification responses to isolated, multispeaker Mandarin stimuli with six glottal periods were contingent on the tones of the stimuli, indicating that the tone responses were not random. Analyses by individual tones revealed the same result (except for Tone 2 stimuli in male speech), further suggesting that tone identification from the brief stimuli exceeded chance. Acoustic analyses showed that dynamic F0 contrasts among the tones were neutralized. Consequently, dynamic F0 information could not have been useful. No context was given; therefore the listeners could not have benefited from any syllable-external information. Familiarity with the speakers' voices is unlikely to have helped because each stimulus was presented only once and the listeners heard each speaker only four times.

The acoustic analyses showed F0 height contrasts between the high-onset tones (Tones 1 and 4) and the low-onset tones (Tones 2 and 3). This finding is consistent with the perceptual results that tones sharing a similar F0 height tend to be confused, and that F0 height identification accuracy exceeded chance. The issue remained, however, of how a particular F0 can be judged as high or low given the speaker variability. The acoustic analyses showed duration, F1 bandwidth (H1-A1), and spectral tilt (H1-A3) were all distinct between the high- and low-onset tones, suggesting that these three acoustic measures covary with F0 to provide information about F0 height. This observation was verified by the correlation analyses. The use of this covariation for perception was further supported by the significant correlation between F0 height identification accuracy and the four acoustic measures (F0, duration, F1 bandwidth, and spectral tilt), even though F0 itself was the only statistically significant predictor of tone classification by F0 height in the regression analyses. Nonetheless, these results suggest that the covariation among F0, duration, F1 bandwidth and spectral tilt was exploited for F0 height estimation, which served as the basis

for tone identification from these brief stimuli without context, dynamic F0 information, or prior exposure to the speakers' voices.

The speaker gender effect found in the perceptual accuracy measure and in the acoustic measures of F0, duration, F1 bandwidth, and spectral tilt also suggest that gender detection may be implicated in the tone identification process. The differences between female and male speakers in F0 (Peterson and Barney, 1952) and voice quality measures (Hanson, 1997; Hanson and Chuang, 1999) have been reported in the literature. However, F0 height information alone will not be useful because an F0 value could indicate a low tone for a female speaker or a high tone for a male speaker, as the current acoustic results showed (Fig. 2). In contrast, the voice quality measures consistently distinguished the female from the male voice. Since gender differences have been shown with these measures (Hanson, 1997; Hanson and Chuang, 1999), the listeners could have used these voice quality differences to determine whether a speaker was female or male, and then evaluated F0 height based on the gender information.

Honorof and Whalen (2005) offered a similar, gender-based interpretation of their finding that English-speaking listeners were able to locate an F0 reliably within a speaker's F0 range without context or prior exposure to a speaker's voice. In particular, F0 location judgments were positively correlated with the female and male F0 range values. This result suggests that F0 height may be estimated based on templates of F0 ranges for female and male speakers ("population tessitures") stored in a listener's memory. As noted in the introduction, a similar idea has also been proposed by Deutsch and co-workers (Deutsch, 1991; Deutsch, *et al.*, 1999, 2004; Deutsch, *et al.*, 1990). In particular, listeners acquire pitch class templates of prevalent speaking F0s from experience with the speakers in their linguistic community, and these templates are used for both speech production and speech perception.

For this gender-based strategy to work, the speaking F0 range in the linguistic community has to be fairly constrained. The F0 difference among the tones must also exceed the F0 variability among the speakers of a particular gender. Dolson (1994) reported that the average speaking F0 is within plus/minus three semitones for either gender within a linguistic community. The acoustic data from the current study also provide some support for this strategy. In particular, despite the considerable number of speakers used (16 females and 16 males), the error bars in Fig. 2 show that the F0 height for a particular tone does not vary too much within a gender group, particularly for the female speakers. In addition, the high-low tonal distinction is fairly well maintained within either gender group, although more prominently for the female speakers. With these reasonably constrained and distinct F0 ranges for both gender groups, gender detection may be used as a first pass before F0 height can be evaluated based on the gender-specific F0 templates.

However, this account does not specify how gender is actually detected. The acoustic data in this study suggest voice quality measures such as F1 bandwidth and spectral tilt may be implicated (Hanson, 1997; Hanson and Chuang,

1999). The covariation between these measures and F0 as well as the significant correlations between these voice quality measures and F0 height detection accuracy also support the perceptual role of the voice quality measures in gender detection. However, since the listeners in the current study were not asked to identify the gender of the speakers, it is not known if they were indeed able to identify gender from the stimuli. Another potential source of gender identification is formant frequency cues to vocal tract length. Bachorowski and Owren (1999) showed speaker gender can be classified acoustically with 92% accuracy from a database of a vowel produced by 125 speakers. Ingemann (1968) also showed that speaker gender can be identified with 75% accuracy from isolated [s] tokens produced by 14 speakers. The stimuli used in the current study happened to include both the fricative and some vowel information. The listeners could have used both the glottal source characteristics (indicated by the voice quality measures) and vocal tract filter information (represented by resonant frequencies from the fricative or vowel) to come up with a gender decision before using the gender information to make F0 height judgments. Further research is needed to test this hypothesis, but the available evidence seems to suggest that gender detection is involved.

The findings from the current study can also be compared to the literature on tone identification from incomplete acoustic input. The stimuli in the current study are generally shorter than the shortest fragment used in previous gating studies (Tseng, 1981; Whalen and Xu, 1992; Wu and Shu, 2003; Yang, 1992). Inspection of Whalen and Xu's (1992) data showed Tones 1 and 4 can be identified with 75% or higher accuracy with 80 ms of input, but Tones 2 and 3 can only be identified at 50% and 40% correct with 100 ms of input. However, if the high- versus low-onset classification is applied instead of specific tones, the data indicate that the high-low distinction can be perceived right at the first gate (40 ms). Wu and Shu (2003) found that a greater amount of acoustic input is needed for Tone 2 identification compared to the other tones (Tone 1: 156 ms; Tone 2: 178 ms; Tone 3: 151 ms; Tone 4: 148 ms). When the absolute values were converted to a percentage of the entire stimulus, however, Tone 2 no longer needed the longest input for identification. Rather, there was no difference among the four tones except for Tone 3 (Tone 1: 56%; Tone 2: 59%; Tone 3: 42%; Tone 4: 60%). These percentages are consistent with Tseng (1981) and Yang (1992), who showed that isolated Mandarin tones can be identified on the basis of the first half of a vowel/syllable. By most accounts (e.g., Whalen and Xu, 1992), half a vowel/syllable is the point where the F0 contours of the four tones become acoustically distinct. The conclusion to be drawn from existing evidence, then, is that the identification of specific tones requires F0 contour information, which requires about half of a vowel/syllable. Tone classification based on F0 height, in contrast, needs very little input. The current study further showed that F0 height can be detected from multispeaker stimuli when no context, dynamic F0, or prior exposure to speaker voice is available.

More broadly, these findings have implications for the efficiency of processing prosodic information in spoken lan-

guage comprehension. In particular, the processing of tonal information is normally considered to take place later than the processing of segmental information due to the temporally distributed nature of tones (Cutler and Chen, 1997). Previous studies showed tones presented in context can be identified quite early (Lee, 2000; Xu, 1994, 2004). The current findings further suggest that multispeaker, isolated tones can be processed in terms of their F0 height information fairly early as well. These findings are consistent with gating studies showing that lexical stress and lexical pitch accent can be identified with the input of one syllable (Cutler and Otake, 1999; van Heuven, 1988, cited in Cutler and Donselaar, 2001; Cutler, et al., 2007). Since stress and pitch accent contrasts are relative and presumably involve the comparison of at least two syllables, it seems counterintuitive to be able to detect stress or pitch accent based only on one syllable. Nonetheless, the stimuli in these studies were recorded by one speaker and were presented with a carrier phrase. Familiarity with speaker voice and the presence of context could have contributed to the evaluation of the acoustic cues available in the stimuli. It will be of interest to evaluate whether listeners can identify these prosodic contrasts in isolated stimuli produced by multiple speakers in the same way as they processed brief lexical tone stimuli in the current study.

ACKNOWLEDGMENTS

I am grateful to Allard Jongman and two anonymous reviewers for their helpful comments. I also thank Z. S. Bond for discussions, Ya-Ting Shih for assistance in administering the perception experiment, and Ning Zhou, Alex Sergeev, Anne Marie Christy, and Gayatri Ram for assistance in data processing. This research was supported in part by a faculty development fund from the School of Hearing, Speech and Language Sciences at Ohio University.

- Abel, S. M. (1972). "Duration discrimination of noise and tone bursts," *J. Acoust. Soc. Am.* **51**, 1219–1223.
- Abramson, A. S. (1972). "Tonal experiments with whispered Thai," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, edited by A. Valdman (Mouton, The Hague), pp. 31–44.
- Abramson, A. S. (1978). "Static and dynamic acoustics in distinctive tones," *Lang Speech* **21**, 319–325.
- Andruski, J. E. (2006). "Tone clarity in mixed pitch/phonation-type tones," *J. Phonetics* **34**, 388–404.
- Andruski, J. E., and Ratliff, M. (2000). "Phonation types in production of phonological tones: The case of Green Mong," *J. Int. Phonetic Assoc.* **30**, 37–61.
- Bachorowski, J.-A., and Owren, M. J. (1999). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," *J. Acoust. Soc. Am.* **106**, 1054–1063.
- Blicher, D. L., Diehl, R. L., and Cohen, L. B. (1990). "Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: Evidence of auditory enhancement," *J. Phonetics* **18**, 37–49.
- Childers, D. G., and Wu, K. (1991). "Gender recognition from speech. Part II: Fine analysis," *J. Acoust. Soc. Am.* **90**, 1841–1856.
- Cohen, B. H. (1996). *Explaining Psychological Statistics* (Brooks/Cole, Pacific Grove).
- Cutler, A., and Chen, H.-C. (1997). "Lexical tone in Cantonese spoken word processing," *Percept. Psychophys.* **59**, 165–179.
- Cutler, A., and Donselaar, W. (2001). "Voornaam is not a homophone: Lexical prosody and lexical access in Dutch," *Lang Speech* **44**, 171–195.
- Cutler, A., and Otake, T. (1999). "Pitch accent in spoken-word recognition in Japanese," *J. Acoust. Soc. Am.* **105**, 1977–1988.
- Cutler, A., Wales, R., Cooper, N., and Janssen, J. (2007). "Dutch listeners' use of suprasegmental cues to English stress," *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1913–1916.
- Deutsch, D. (1991). "The tritone paradox: An influence of language on music perception," *Music Percept.* **8**, 335–347.
- Deutsch, D., Henthorn, T., and Dolson, M. (1999). "Absolute pitch is demonstrated in speakers of tone languages," *J. Acoust. Soc. Am.* **106**, 2267.
- Deutsch, D., Henthorn, T., and Dolson, M. (2004). "Absolute pitch, speech, and tone language: Some experiments and a proposed framework," *Music Percept.* **21**, 339–356.
- Deutsch, D., North, T., and Ray, L. (1990). "The tritone paradox: Correlate with the listener's vocal range for speech," *Music Percept.* **7**, 371–384.
- Dolson, M. (1994). "The pitch of speech as a function of linguistic community," *Music Percept.* **11**, 321–331.
- Dooley, G. J., and Moore, B. C. J. (1988). "Duration discrimination of steady and gliding tones: A new method for estimating sensitivity to rate of change," *J. Acoust. Soc. Am.* **84**, 1332–1337.
- Fox, R. A., and Qi, Y.-Y. (1990). "Context effects in the perception of lexical tones," *J. Chin. Linguist.* **18**, 261–284.
- Fox, R. A., and Unkefer, J. (1985). "The effect of lexical status on the perception of tone," *J. Chin. Linguist.* **13**, 69–90.
- Gandour, J. (1983). "Tone perception in Far Eastern languages," *J. Phonetics* **11**, 149–175.
- Gandour, J. T., and Harshman, R. A. (1978). "Crosslanguage differences in tone perception: A multidimensional scaling investigation," *Lang Speech* **22**, 1–33.
- Gelfand, S. A. (1998). *Hearing: An Introduction to Psychological and Physiological Acoustics* (Dekker, New York).
- Gottfried, T. L., and Suiter, T. L. (1997). "Effects of linguistic experience on the identification of Mandarin Chinese vowels and tones," *J. Phonetics* **25**, 207–231.
- Greenberg, S., and Zee, E. (1979). "On the perception of contour tones," *UCLA Working Papers in Phonetics* **45**, 150–164.
- Grosjean, F. (1996). "Gating," *Lang. Cognit. Processes* **11**, 597–604.
- Hanson, H. M. (1997). "Glottal characteristics of female speakers: Acoustic correlates," *J. Acoust. Soc. Am.* **101**, 466–481.
- Hanson, H. M., and Chuang, E. S. (1999). "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.* **106**, 1064–1077.
- Honorof, D. N., and Whalen, D. H. (2005). "Perception of pitch location within a speaker's F0 range," *J. Acoust. Soc. Am.* **117**, 2193–2200.
- Howell, D. C. (1999). *Fundamental Statistics for the Behavioral Sciences* (Brooks/Cole, Pacific Grove, CA).
- Howie, J. M. (1976). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge).
- Huffman, M. (1987). "Measures of phonation type in Hmong," *J. Acoust. Soc. Am.* **81**, 495.
- Ingemann, F. (1968). "Identification of the speaker's sex from voiceless fricatives," *J. Acoust. Soc. Am.* **44**, 1142–1144.
- Johnson, K. A. (2005). "Speaker normalization in speech perception," *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 363–389.
- Ladefoged, P. (1996). *The Sounds of the World's Languages* (Blackwell, Malden, MA).
- Lass, N. J., Mertz, P. J., and Kimmel, K. L. (1978). "Effect of temporal speech alterations on speaker race and sex identifications," *Lang Speech* **21**, 279–290.
- Leather, J. (1983). "Speaker normalization in perception of lexical tone," *J. Phonetics* **11**, 373–382.
- Lee, C.-Y. (2000). "Lexical tone in spoken word recognition: A view from Mandarin Chinese," thesis, Brown University, Providence, RI.
- Lee, C.-Y., Tao, L., and Bond, Z. S. (2008). "Identification of acoustically modified Mandarin tones by native listeners," *J. Phonetics* **36**, 537–563.
- 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.
- Liu, S., and Samuel, A. G. (2004). "Perception of Mandarin lexical tones when F0 information is neutralized," *Lang Speech* **47**, 109–138.
- Massaro, D. W., Cohen, M. M., and Tseng, C.-Y. (1985). "The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese," *J. Chin. Linguist.* **13**, 267–289.
- Mertus, J. A. (2000). *The Brown Lab Interactive Speech System* (Brown University, Providence, RI).
- Moore, C. B., and Jongman, A. (1997). "Speaker normalization in the perception of Mandarin Chinese tones," *J. Acoust. Soc. Am.* **102**, 1864–1877.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech

- perception," *Percept. Psychophys.* **60**, 355–376.
- Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993). "Episodic encoding of voice attributes and recognition memory for spoken words," *J. Exp. Psychol. Learn. Mem. Cogn.* **19**, 309–328.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Sinnott, J. M., Owren, M. J., and Peterson, M. R. (1987). "Auditory duration discrimination in Old World monkeys (*Macaca. Cercopithecus*) and humans," *J. Acoust. Soc. Am.* **82**, 465–470.
- Stevens, K. N. (1998). *Acoustic Phonetics* (MIT, Cambridge, MA).
- Stevens, K. N., Li, Z., Lee, C.-Y., and Keyser, J. (2004). "A note on Mandarin fricatives and enhancement," in *From Traditional Phonology to Modern Speech Processing*, edited by G. Fant, H. Fujisaki, J. Cao, and Y. Xu (Foreign Language Teaching and Research, Beijing).
- Swerts, M., and Veldhuis, R. (2001). "The effect of speech melody on voice quality," *Speech Commun.* **22**, 297–303.
- Tabachnick, B. G., and Fidell, L. S. (2001). *Using Multivariate Statistics*, 4th ed. (Allyn and Bacon, Boston).
- Tseng, C.-Y. (1981). "An acoustic phonetic study on tones in Mandarin Chinese," Ph.D. thesis, Brown University, Providence, RI.
- Heuven, V. J. (1988). "Effects of stress and accent on the human recognition of word fragments in spoken context: Gating and shadowing," *Proceedings of Speech '88, seventh FASE Symposium*, Edingburgh, pp. 811–818.
- Wang, H. (1986). *A Frequency Dictionary of Modern Chinese* (Beijing Language Institute, Beijing).
- Whalen, D. H., and Xu, Y. (1992). "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica* **49**, 25–47.
- Wong, P. C. M., and Diehl, R. L. (2003). "Perceptual normalization for inter- and intra-talker variation in Cantonese level tones," *J. Speech Lang. Hear. Res.* **46**, 413–421.
- Wu, N., and Shu, H. (2003). "The gating paradigm and spoken word recognition of Chinese," *Acta Psychologica Sinica* **35**, 582–590.
- Xu, Y. (1994). "Production and perception of coarticulated tones," *J. Acoust. Soc. Am.* **95**, 2240–2253.
- Xu, Y. (1997). "Contextual tonal variations in Mandarin," *J. Phonetics* **25**, 61–83.
- Xu, Y. (2004). "Understanding tone from the perspective of production and perception," *Language and Linguistics* **5**, 757–797.
- Yang, S. (1992). "A preliminary study on the perceptual center of tones in Standard Chinese," *Acta Psychologica Sinica* **3**, 247–253.
- Yost, W. A. (2007). *Fundamentals of Hearing: An Introduction* Elsevier, Burlington, MA.