On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements[a]

Jing Shen,[b] Diana Deutsch, and Keith Rayner
Department of Psychology, University of California, San Diego, La Jolla, California 92093

(Received 28 February 2012; revised 6 October 2012; accepted 19 February 2013)

Using the visual world paradigm, the present study investigated on-line processing of fine-grained pitch information prior to lexical access in a tone language; specifically how lexical tone perception of Mandarin Tones 2 and 3 was influenced by the pitch height of the tone at onset, turning point, and offset. Native speakers of Mandarin listened to manipulated tone tokens and selected the corresponding word from four visually presented words (objects in Experiment 1 and characters in Experiment 2) while their eye movements were monitored. The results showed that 87% of ultimate tone judgments were made according to offset pitch height. Tokens with high offset pitch were identified as Tone 2, and low offset pitch as Tone 3. A low turning point pitch served as a pivotal cue for Tone 3, and prompted more eye fixations on Tone 3 items, until the offset pitch directed significantly more fixations to the final tone choice. The findings support the view that lexical tone perception is an incremental process, in which pitch height at critical points serves as an important cue. © *2013 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4795775]

PACS number(s): 43.71.Es, 43.71.Sy [MSV] Pages: 3016–3029

I. INTRODUCTION

The present study employed the visual world paradigm to investigate the hypothesis that fine-grained pitch cues are exploited in an incremental fashion in the pre-lexical perception of isolated lexical tones, in which pitch height at the critical points of the syllable (i.e., onset, turning point, and offset) serves as an important cue. Languages that utilize phonologically contrastive variations in pitch at the lexical level are called tone languages (McCawley, 1978). In tone languages, such as most East Asian and African languages, words with the same vowels and consonants have different meanings depending on the lexical tones in which they are enunciated. For example, the word *ma* in Mandarin means “mother” when spoken in the first tone, “hemp” when spoken in the second tone, “horse” in the third tone, and a reproach in the fourth tone. This contrasts with intonation languages such as English, in which pitch is used to convey emotion and the form of utterance (i.e., statement or question), but is not involved in determining the meaning of individual words. For instance, in Mandarin Chinese, the four lexical tones can be described by their fundamental frequency ($F0$) patterns as high-level, high-rising, low-falling-rising, and high-falling. Other experiments have demonstrated that when the primary cue of pitch is missing or ambiguous, other acoustic cues will be exploited, such as amplitude contour (Whalen and Xu, 1992; Fu and Zeng, 2000) and duration (Blicher *et al.*, 1990).

Research has revealed that native speakers of tone languages utilize several pitch cues to identify lexical tones. A multi-dimensional scaling study indicated that the major dimensions invoked are overall pitch height, direction of pitch change, pitch height at endpoint, and slope of pitch change (Gandour, 1983). A considerable body of research has tested three major pitch cues individually in lexical tone perception, namely, overall pitch height, pitch at the critical points, and slope of pitch change.

A. Overall pitch height

Deutsch *et al.* (2004) demonstrated that native speakers of Mandarin showed a significantly higher pitch consistency compared with native speakers of English when the different groups were asked to enunciate a list of words in their native language on two different days. Further data (Deutsch *et al.*, 2009) collected from two groups of Chinese speakers living in two different isolated linguistic communities showed that the pitch ranges of the subjects’ speech were clustered within each community while differing significantly across communities. These studies provided support for the hypothesis that native speakers of tone languages acquire a mental representation of pitch height through long-term exposure to the pitch range of speech in the speakers’ linguistic community (Deutsch, 1992; Deutsch, *et al.*, 2004; Dolson, 1994). This mental template of pitch height can potentially help tone language speakers exploit overall pitch height as a cue for tone perception, especially if they had only lived in one community and thus only have one consistent pitch template for speech.

Extensive research has been carried out to examine how overall pitch height affects lexical tone perception. First, it was shown in a reaction time study that overall pitch height affected tone judgment in isolated Mandarin syllables (Shen *et al.*, 2011). Furthermore, in paired Mandarin syllables, the

[a] Portions of this work were presented at the 20th International Congress on Acoustics, Sydney, Australia, August 2010, by J. Shen, D. Deutsch, and K. Rayner, “Processing of endpoint pitch in Mandarin tone perception: An eye movement study,” incorporating Proceedings of the 2010 Annual Conference of the Australian Acoustical Society, p. 3666.

[b] Author to whom correspondence should be addressed. Electronic mail: j2shen@ucsd.edu

average and onset $F0$ of one syllable influenced tone judgment for the other syllable (Lin and Wang, 1985; Fox and Qi, 1990). In addition, the effect of overall pitch height has been demonstrated using the paradigm of embedding the syllables in context sentences that are manipulated in pitch, both for Cantonese level tones (Wong and Diehl, 2003; Francis *et al.*, 2006), and for Mandarin contour tones (Leather, 1983; Moore and Jongman, 1997; Huang and Holt, 2009).

While all these studies have suggested a critical role of overall pitch height as a cue for tone perception, this cue could be formed from a variety of sub-cues, such as the average $F0$ of the syllable, the $F0$ range covered by the extreme pitch points, or the entire pitch contour. Further examination of these other pitch cues is therefore needed to investigate this possibility.

B. Endpoint and midpoint pitches

In several studies employing the paradigm of removing part of a tone (Gottfried and Suiter, 1997; Lee *et al.*, 2008; Liu and Samuel, 2004), it was found that native speakers of Mandarin were able to correctly identify most of the tones with which they were presented, even with only a small proportion of the sound signal available. Adopting the paradigm from the study of Strange *et al.* (1983) on vowel identification, Gottfried and Suiter (1997) investigated tone perception by native and non-native speakers of Mandarin using four types of syllable: Intact, center-only (without the initial six and final eight pitch periods), silent-center (only the initial six and final eight pitch periods are available), and onset-only (only the initial six pitch periods are available). Native speakers outperformed non-native speakers overall, especially for the silent-center syllables. The percentage average correct tone identification for the silent-center syllables was approximately 95% for native speakers, compared to 46% for non-native speakers. As a replication and extension of this study, Lee *et al.* (2008) carried out a tone identification study using the same types of stimuli and a larger sample size consisting of 40 native speakers, and obtained a similar result. Using a different tone language, Thai, Zsiga and Nitisaroj (2007) demonstrated in a series of experiments that pitch inflections at the syllable midpoint and offset point successfully categorized Thai tones in perceptual space. The findings from these studies provide evidence that native speakers of tone languages use midpoint and endpoint pitches as perceptual cues for tone identification.

Several studies have investigated the cues of pitch change and the timing of the pitch turning point in Tones 2 (high-rising) and Tone 3 (low-falling-rising). Shen *et al.* (1993) found that the timing of the pitch turning point was an important cue for discriminating between Tone 2 and Tone 3. A turning point that occurred close to the tone onset served as a cue for Tone 2, and one that occurred late in the tone prompted more judgments of Tone 3. Findings by Moore and Jongman (1997) showed that both timing of pitch turning point and pitch difference between onset and turning point influenced perception of isolated tones. The stimuli were more likely to be identified as Tone 2 words when they had an early turning point and a small pitch difference between onset and turning point. Overall, perception of Tone 2 seemed to tolerate more variability, while Tone 3 required a late turning point and a large initial fall in $F0$.

The above studies all focused on the pitch information between onset and turning point for Tones 2 and 3, and the importance of the later parts of the tones has rarely been investigated. Liu and Samuel (2004) provided data on percentage correct identification for tone stimuli that had a portion of the syllable replaced by white noise (either from syllable onset to pitch turning point, or from pitch turning point to syllable offset). Their findings showed that for Tone 2, the information contained in the segment between the turning point and the offset point was critical for identification, while this phenomenon did not hold for Tone 3. Further, it was found that for Tone 3 the first half of the syllable was more important for identification than the second half. While this finding aligns with others in suggesting that some points or segments within the syllable contain particularly salient pitch information for tone identification, it also raises the possibility of combining segments from different tones to create hybrid tones as stimuli for investigating the perception of tones.

C. Slope of pitch change

The existing literature suggests that, along with pitch height, slope of pitch change is a perceptual cue used by native speakers for tone identification (Gandour, 1983; Massaro *et al.*, 1985; Chandrasekaran *et al.*, 2007). Research on lexical tone perception has consistently shown that native speakers perceive pitch glides categorically, with pitch patterns across the categorical boundary perceived as different tones, while those within the boundary perceived as the same tone (Francis *et al.*, 2003; Xu *et al.*, 2006; Halle *et al.*, 2004). For example, modeling Mandarin high level and high rising tones, Xu *et al.* (2006) resynthesized a series of speech sound tokens carrying pitch glides with the same offset pitch height but different linear rising slopes. They found that native speakers identified sound tokens with slopes of pitch change larger than approximately 0.047 Hz/ms as a high rising tone, while those with shallow slopes as a high level tone. Similar results were reported in two other studies using different tone languages and subject groups: Cantonese rising and falling tones by Francis *et al.* (2003), and Taiwan Mandarin level, rising, and falling tones by Halle *et al.* (2004).

In research on Mandarin tone perception, a continuum based on high level and high rising tones (Tones 1 and 2) has been frequently used for creating stimuli in identification and discrimination tasks (Chan *et al.*, 1975; Wang, 1976; Xu *et al.*, 2006). However, there has not been any investigation of the rising slopes between pitch turning point and offset in Tones 2 and 3 to determine whether a steep slope of pitch change alone can prompt the judgment of Tone 2 and the lack of a steep slope can serve as a cue for Tone 3.

The existing tone perception literature has mostly employed paradigms in which subjects identified and/or discriminated between tones after the tokens were presented. By manipulating sound to control for different acoustic cues, this method can provide information regarding the acoustic cues that influence tone perception by native speakers. However, because these cues usually co-vary (e.g., pitch height at critical points can co-vary with the slope of pitch change), measuring only the final tone judgment has made it difficult to tease apart the effects of several different factors. For example, the experiment performed by Moore and Jongman (1997) involved three pitch cues: (1) Pitch height at onset (with pitch height at turning point kept consistent), (2) pitch difference between onset and turning point, and (3) the slope of pitch change between the onset and turning point. Changing one of these cues in such cases affects the other two, making it difficult to know which cue is most influential in the final tone identification.

Another intriguing question is: How are these pitch cues utilized by native speakers of tone languages in perceiving lexical tones? Are these pitch cues used in an incremental way and change the cumulative evidence for identifying the tone as the speech sound unfolds? Or are they processed in a holistic manner so as to make the judgment after the entire syllable has been presented? The traditional paradigm that only records off-line responses cannot provide an answer to this question.

Taken together, the problem and the question demand an on-line paradigm that can reflect any instant change of tone judgment before the syllable ends. The visual world paradigm is a good choice for tackling this problem. Because of the incremental fashion in which listeners recognize spoken words and comprehend speech, this eye tracking method has been used to study on-line speech perception (Rayner and Clifton, 2009). In a typical visual world study, the subject follows instructions to either complete certain tasks with one of a small set of objects presented in a visual workspace while receiving auditory stimuli (see Tanenhaus *et al.*, 1995) or simply hears spoken sentences and views the objects on a screen without performing any explicit task (see Huetting and Altmann, 2007). There are two basic components of eye movements in such a task. Fixation refers to the period of time when the eyes remain still at a location and extract visual information. Saccades are the movements themselves, during which there is no new information acquired because vision is suppressed (Matin, 1974). In this task, either eye fixation or covert attention are to the same location (i.e., during a fixation), or attention precedes the eyes to the next saccade location (Rayner, 2009). The subjects are pre-exposed to a set of visual objects before the sound stimuli are presented, and then make saccades and fixations on these objects while the speech sound is playing. When the data are aggregated across a large number of trials, there are, overall, a high proportion of fixations (POF) on the object that corresponds to the auditory input at a certain time point. The timing and pattern of fixation to potential referents can be used to draw inferences about perception and comprehension, as the amount of fixation on a visual object is considered to be a reflection of the amount of neural activation associated with the word (Tanenhaus *et al.*, 2000). This paradigm has lately been adapted to study sub-phonetic processing in speech perception, so as to investigate how fine-grained acoustic differences can be used to map speech sounds into phonemes and words (McMurray *et al.*, 2002; Dahan *et al.*, 2001). McMurray *et al.* (2002) used the visual world paradigm to demonstrate the gradient effect of voice onset time (VOT) on lexical activation. They used pairs of words that only have initial consonants which differed in VOT (e.g., /b/ vs /p/) and synthesized a series of words with a nine-step VOT continuum. The subjects were instructed to click on the objects they heard while their eye movements were monitored. It was found that as VOT approached the categorical boundary, even though the subjects eventually selected the target objects, the fixations on the lexical competitor that differed in voicing increased. Dahan *et al.* (2001) created subcategorical mismatches by using cross-spliced words that contained mismatched acoustic information between vowels and consonants. The eye fixation data showed a significant effect on on-line lexical activation, coming from the fine-grained co-articulatory information at the onset of the target word. These studies demonstrated the effectiveness of the visual world paradigm for examining time-sensitive phonetic and lexical processing in response to the fine-grained acoustic information in the speech sound.

Although this paradigm has been used intensively in studying spoken word recognition, only a few studies have so far examined on-line processing of lexical tones. Malins and Joanisse (2010) employed the visual world paradigm to examine how tonal versus segmental information influence spoken word recognition in Mandarin Chinese. By comparing the time course of viewing items that share the same segmental information versus items that share the same tonal information, they concluded that segmental and tonal information are accessed concurrently, and play a comparable role in Mandarin word recognition. Also using the visual world paradigm, Speer and Xu (2005) examined the effect of lexical tones on word recognition during the comprehension of continuous speech. Listeners' eye movements were monitored as they listened to Mandarin sentences that contained ambiguous word sequences resulting from the operation of the third tone sandhi rule. Their findings suggest a very early use of tonal information in identifying the words. While extending the spoken word recognition literature by taking tonal information into account, these studies focused on the comparison of tonal and segmental processing as parallel mechanisms in spoken word recognition. However, no study has so far investigated on-line processing of pitch-tone mapping by manipulating the acoustic cues in speech sounds.

The motivation of the present study was two-fold. First, building on the literature concerning pitch cues for tone perception, we examined how native speakers utilize the two acoustic cues (pitch height at critical points, and slopes of pitch change) for discriminating Mandarin Tones 2 and 3. Along with the manipulation of $F0$ at syllable onset, turning point, and offset, the eye tracking paradigm enabled us to determine native speakers' response to acoustic information at specific time points within the syllable; this provided valuable information concerning how each of these pitch cues influenced tone judgments.

Second, the study extended research on spoken word recognition to investigate the processing of fine-grained acoustic information prior to lexical access. Lexical tone serves this goal nicely in the sense that the lexical meaning of syllables in a tone language differs only depending on pitch information given that vowels and consonants are held constant. As the visual world paradigm presents data on a timescale of milliseconds, it satisfies the need to examine on-line perceptual responses to instantaneous pitch changes in a tone as the sound unfolds. In the present study, the processing of the pitch information was revealed in a dynamic way by examining the time pattern of attention (i.e., measured by eye fixations) directed to potential referents, and how this changed with the unfolding of the speech sound.

Both experiments had the same design. The factor of offset pitch had four levels: Original high, ambiguous high, original low, and ambiguous low. The four conditions were termed *High Tone 2* condition (with original high offset pitch), *Low Tone 2* condition (with ambiguous high offset pitch), *High Tone 3* condition (with ambiguous low offset pitch), and *Low Tone 3* condition (with original low offset pitch). In all the conditions, the onset and turning point pitch were set to have the same low pitch values representing those pitch cues in Tone 3.

The present study was motivated by two hypotheses. First, the data of Gottfried and Suiter (1997) and Lee *et al.* (2008) showed that native speakers were able to accurately identify tones with only onset and offset pitch information. It is predicted on this hypothesis that low onset pitch should give listeners an initial impression of Mandarin Tone 3. After the entire syllable is heard, a low offset pitch should confirm the earlier choice of Tone 3 while a high offset pitch should change most of the final judgments to Tone 2, given the finding that the second half of the tone is more important for identifying Tone 2 (Liu and Samuel, 2004).

Second, because the offset pitch values of High Tone 2 and Low Tone 3 were taken from the particular speaker's pitch of speech, these two conditions served as the original conditions. The two ambiguous conditions were created by increasing the offset pitch in the Low Tone 3 condition by one semitone, and by decreasing that of the High Tone 2 condition by one semitone. Shen *et al.* (2011) found that a difference of merely 1.5 semitones in overall pitch height significantly influenced how Tone 3 was identified. If the listeners in the present study responded to small differences in offset pitch height in the same way, the final tone judgment should differ for the original and ambiguous conditions.

Due to the close correspondence between orthographic and phonological information in alphabetic languages, studies using the visual world paradigm typically use pictures of objects as visual stimuli (e.g., Tanenhaus *et al.*, 1995; Huetting and Altmann, 2007). Experiment 1 of the present study followed this paradigm to obtain results comparable to those in the existing literature. Additionally, the stimuli in the present study were in Chinese, which are logographic in nature with highly arbitrary symbol-sound correspondences (Wang, 1973; Zhou *et al.*, 1999). As the written form of words (i.e., characters) does not correspond to lexical tones in Chinese, characters could then serve as visual stimuli in the visual world paradigm. Although Chinese characters and pictures of objects have both been used in eye tracking paradigms to study tone perception (Speer and Xu, 2005; Malins and Joanisse, 2010), no study so far has employed these two paradigms to investigate the same questions. Experiment 2 of the present study aimed at testing the paradigm of using characters rather than pictures.

II. EXPERIMENT 1

A. Method

1. Subjects

Twenty-four native Mandarin speakers (11 males, 13 females) were recruited from international and visiting students and scholars at the University of California, San Diego. The average age of the subjects was 26.8 (standard deviation = 4.1). All subjects were from Mainland China and had been living in the United States for a mean of 7.9 months and a range of 1 to 20 months. None of them had more than 5 years of musical training, and none had any recent musical training within the past 10 years. No subject reported any hearing or speech disorders, and they all had normal or corrected to normal vision. Informed consent was obtained prior to the experiment and the subjects were paid for their participation. Three additional subjects were tested but their data were excluded from all the analyses due to excessive eye blinking (over 30% of trials).

2. Stimuli and instrumentation

Speech tokens of eight syllables, which have meanings for both Tones 2 and 3 in Mandarin, were recorded by a female native speaker in a sound attenuated booth using a Zoom H2 digital audio recorder (Zoom Corp., Tokyo), and saved as WAV files at a sampling rate of 44.1 K and a 16 bit resolution. (See Table I for Pinyin of these syllables and name of the objects.) None of these syllables had fricative consonants. The voiced portion of the sound tokens, which carried the pitch information, began no later than 30 ms into the syllable. The eye movement data corresponding to the unvoiced portion was discarded from the analysis, so the term "tone onset" is used hereafter to refer to the onset of the voiced portion of the syllable. All the syllables ended with either a vowel or a nasal consonant.

Using Praat software (Boersma and Weenink, 2008), one pitch value of the speech stimulus was extracted every 10 ms. The onset, offset, and turning point $F0$s were set as the predefined values (see Table II). All the other $F0$ values within the syllable were estimated by a parabolic interpolation method using Praat in order to retain the naturalness of the speech sound (Xu and Sun, 2001, see Fig. 1). The tokens were resynthesized from the natural speech template using the method of Time-Domain Pitch-Synchronous Overlap-and-Add (Moulines and Charpentier, 1990), keeping the formant patterns and amplitudes constant, and normalizing the duration to 500 ms (with pitch turning point at 200 ms).

These speech tokens were first identified by six native speakers of Mandarin to ensure that they sounded natural enough to be identified as Mandarin Tones 2 and 3. These subjects identified tones with high offset pitches as Tone 2 in over 80% of trials, and stimuli with low offset pitches were

TABLE I. Sound stimuli used in Experiment 1 (visual stimuli: pictures of objects).

	Pinyin	Object
Tone 2 stimuli	/bi/	nose
	/lei/	thunder
	/lian/	curtain
	/ling/	bell
	/liu/	stream
	/wei/	go (the game)
	/yan/	rock
	/yu/	fish
Tone 3 stimuli (same syllables as Tone 2 stimuli)	/bi/	pen
	/lei/	bud
	/lian/	mask
	/ling/	collar
	/liu/	willow
	/wei/	tail
	/yan/	eye
	/yu/	rain
Tone 1 (distractors)	/che/	car
	/dao/	knife
	/deng/	lamp
	/ding/	nail
	/gou/	hook
	/ji/	chicken
	/shu/	book
	/shua/	brush
	/ti/	ladder
	/zhong/	clock
Tone 4 (distractors)	/chi/	wing
	/dou/	bean
	/jian/	arrow
	/jing/	mirror
	/ju/	saw
	/mao/	hat
	/pao/	cannon
	/tu/	rabbit
	/xiang/	elephant
	/xie/	crab

identified as Tone 3 in over 90% of trials. No subject reported any unnaturalness in these stimuli.

Each sound token was presented twice during the experiment. To prevent subjects' anticipation of the experiment stimuli, the same number of filler trials was created, in which speech tokens of ten Tone 1 and ten Tone 4 words were presented. None of the Tone 1 and Tone 4 words shared the same syllable as any of the Tone 2 and Tone 3 words and none of the Tone 1 words shared the same syllable with any of the Tone 4 words (see Table I for Pinyin of these syllables and names of objects). All the words were normalized to the same peak intensity using the software Peak Pro Version 5.2 (Bias, Petaluma, CA). They were then transferred to a Dell Precision 390 desktop computer (Dell Inc., Round Rock, TX) and were presented at ~75 dB sound pressure level through a BOSE Companion II speaker system (Bose Corp., Framingham, MA). The visual stimuli were black line-drawings of objects that corresponded to these syllables. They were all resized to 200 × 200 pixels. All the visual stimuli were presented in black [red-green-blue (RGB) code 0-0-0] on a gray (RGB code 135-135-135) background on a 19-in. ViewSonic LCD monitor (ViewSonic Corp., Walnut, CA). The eye movement data were collected using a SR EyeLink1000 eye tracker (SR Ltd., Canada) and a Dell Precision 390 desktop computer.

TABLE II. Pitch of onset, turning point, and offset in four conditions (in Hz).

Condition	Onset pitch height	Turning point pitch height	Offset pitch height
Low Tone 3	184.9	138.6	155.6
High Tone 3	184.9	138.6	164.8
Low Tone 2	184.9	138.6	246.9
High Tone 2	184.9	138.6	261.6

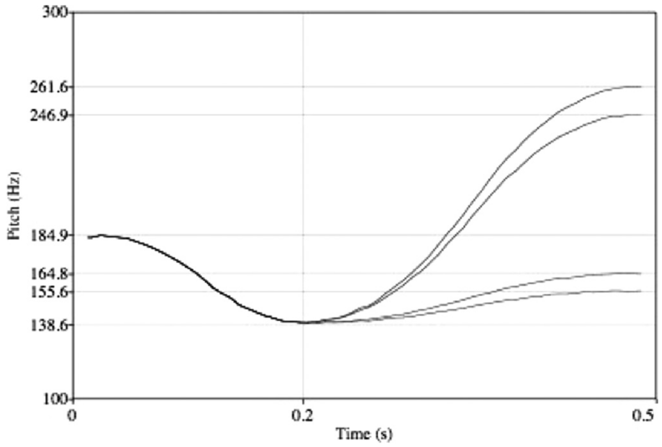

FIG. 1. Pitch patterns of tone stimuli in four conditions (offset pitch heights from top to bottom: *High Tone 2* condition, *Low Tone 2* condition, *High Tone 3* condition, *Low Tone 3* condition).

3. Procedure

The experimental software was developed using MATLAB 7.8.0 (Mathworks, 2009). The subjects were tested individually. After arriving at the lab, the eye tracker in remote setup was calibrated with the standard 9-point calibration procedure. Viewing was binocular, but eye movements were recorded from the right eye only. The sampling rate of the eye tracker was set as 500 Hz. The order of trials was randomized for each subject.

The experiment began with two blocks of training trials to familiarize the subject with the names of the pictures. In the first block, each picture was displayed alone once together with its printed name. In the second block, a printed picture name was displayed along with four candidate pictures. Subjects were required to click on the correct picture to advance to the next trial.

Ten practice trials that were identical to the test trials were given prior to beginning the formal experiment. Each trial started with a standard drift correction procedure, which measures how much the difference between a participant's fixation and a central point "drifts" over a short time period. Drift can occur because of factors such as fatigue and changes in body (head) position. Then a small black box (20 × 20 pixels) appeared at the center of the screen. The subjects were instructed to click on the box to activate the trial. Once the box was clicked on, four pictures, which were each 200 pixels (about 5° of visual angle) in height and width, were displayed on the centers of the four quadrants of the screen. On each trial, the four pictures consisted of a pair of two pictures corresponding to Tones 2 and 3 objects, one Tone 1 object and one Tone 4 object. The Tone 2 and 3 objects shared the same syllable. A set of pictures of objects associated with the Tone 1 and Tone 4 words that were used in filler trials were presented as Tone 1 and Tone 4 objects. None of the Tone 1 and Tone 4 objects shared the same syllable as any of the Tone 2 and Tone 3 objects, and none of the Tone 1 objects shared the same syllable with any of the Tone 4 objects. (See Table I for Pinyin of these syllables and names of the objects, and Fig. 2 for an example of the visual stimuli.) The locations of these pictures on the display were randomized. The sound stimulus, which consisted of sound token preceded by a 700 ms prompt sentence of "Now click on __", was played simultaneously with the display of the pictures. The subjects were asked to complete the task following the instruction given by the sound and to make their best decision with no time constraints.

B. Results and discussion

1. Tone identification data

Analysis of the tone identification data showed that the subjects had an overall rate of correct response of 84.7% at the end of the syllables. The breakdown of the correct response rate for the four conditions was 99.7% (selecting Tone 3 in the *Low Tone 3* condition); 98.6% (selecting Tone 3 in the *High Tone 3* condition); 67.1% (selecting Tone 2 in the *Low Tone 2* condition); and 74.2% (selecting Tone 2 in the *High Tone 2* condition). For further analyses, the selected objects are termed "targets"; the objects associated with the same syllables as the targets but different tones are termed "competitors"; those Tone 1 and 4 objects that were presented on the same screen are termed "distractors."

To make these proportional data more suitable for analysis of variance (ANOVA), a rationalized arcsine transform (Studebaker, 1985) was performed to convert the proportions into R scores before running them through the ANOVAs. In the two low offset conditions, subjects ultimately selected objects associated with Tone 3 more often than those associated with Tone 2, but in the two high offset conditions they selected more Tone 2 than Tone 3 objects. These differences were all significant [$ps < 0.01$], suggesting that the manipulation of offset pitch height and rising slope influenced tone identification. Furthermore, both tones were more frequently selected in the original conditions: Tone 3 was more frequently selected in the *Low Tone 3* condition compared with the *High Tone 3* condition [$t(23) = 2.04$, $p = 0.05$], and Tone 2 was selected more often in the *High Tone 2* condition compared with the *Low Tone 2* condition [$t(23) = 1.68$, $p = 0.1$]. This finding indicates native speakers can exploit subtle pitch cues of one semitone difference in making tone judgments.

2. Temporal window analysis

The eye tracker recorded the subjects' fixation positions (i.e., coordinates on the screen) at 2 ms intervals. These coordinate data were then converted to one data point every 2 ms to show on which one of the four objects the eyes were fixated at that moment. To account for errors of calibration and drift in the eye tracker, the four locations containing the objects were defined as four quadrants that were 640 pixels wide and 512 pixels high each.

Because the present study intended to examine the fixation data in those trials that the subjects made a tone judgment that was consistent with the pitch cue at the tone offset (i.e., selecting targets instead of competitors or distractors), those trials in which the subjects selected objects other than the targets (15.3%) were excluded from statistical analysis of the eye fixation data. The trials that contained blinks (8.3%) during playing of the sound tokens were also excluded from the analysis of the eye fixation data.

The number of fixations was counted in every 20 ms time window, which gives 10 data counts per window to produce the total counts of eye fixations on the target, competitor, and the 2 distractors in each window. For each 20 ms window, the POF data were then derived from the count data by dividing each of the 4 fixation count data (i.e., counts of fixation at target, counts of fixation at competitor, and counts of fixation at the distractors) by the sum of counts of 4 possible locations. The POF data were then collapsed over subjects and items to create 4 averaged trajectory graphs in 20 ms time windows (see Fig. 3), for each of the four conditions.

Prior eye tracking research suggests that, in the visual world paradigm, the time to program and execute a saccadic eye movement is approximately 150 to 200 ms or less (Rayner, 1998, 2009). To test the significance of the difference between fixation on targets and competitors while avoiding too many ANOVAs for each condition, the POF data were collapsed within four 150 ms time windows. The first window was taken as 301 to 450 ms from tone onset.

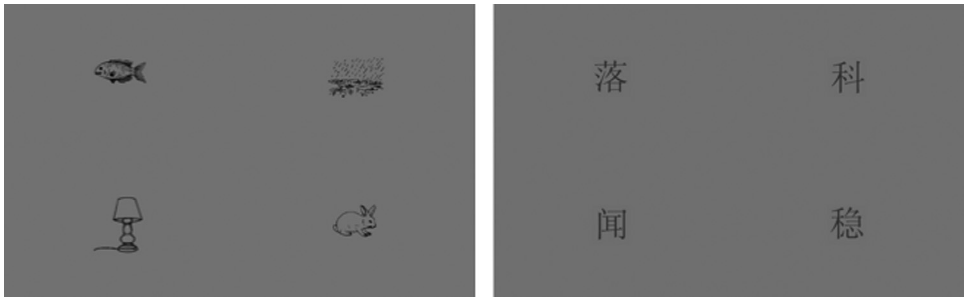

FIG. 2. Examples of the visual stimuli in the two experiments. Left panel: Four pictures (starting top-left clockwise) /yu2/, /yu3/, /tu4/, /deng1/ in Experiment 1. Right panel: Four characters (starting top-left clockwise) /luo4/, /ke1/, /wen3/, /wen2/ in Experiment 2.

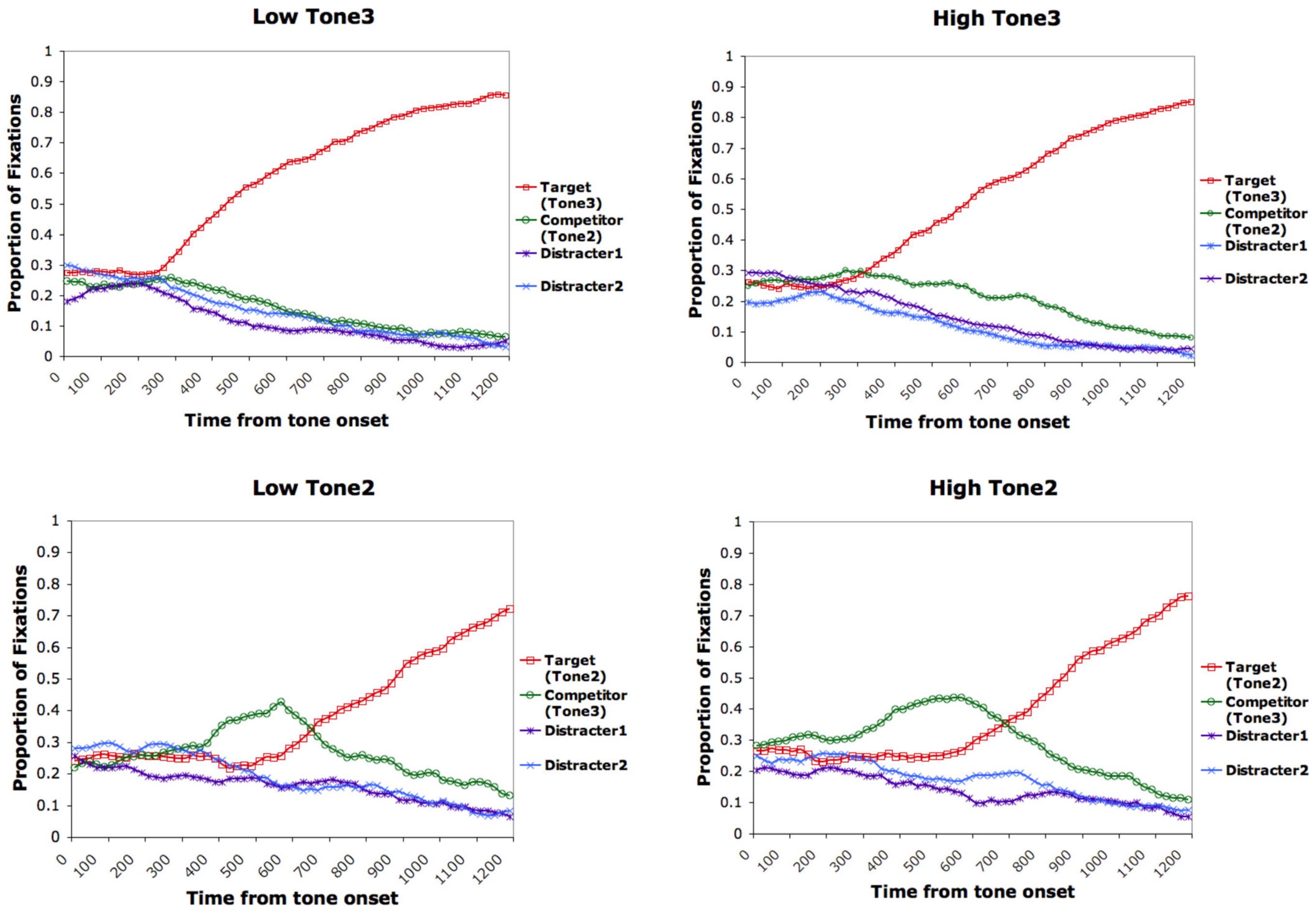

FIG. 3. (Color online) Experiment 1 POF curves in 20 ms interval (time scale in milliseconds).

The fixation data in this window should reflect the influence of the sound from tone onset (i.e., 1 to 100 ms into the tone) to pitch turning point (i.e., until 250 ms into the tone) because of the 200 ms time delay for saccade planning and execution. Any effect from the pitch information up to 100 ms into the tone should fully reveal itself by the beginning of the first window. The last window was defined as 751 to 900 ms from tone onset. Considering the approximately 200 ms saccade planning and execution time, if pitch information at syllable offset serves as a cue for identifying the tones, the fixations should be significantly more likely to be on the target compared to the competitor by the beginning of the last window. The time spans of the 4 windows were: 301 to 450 ms; 451 to 600 ms; 601 to 750 ms; 751 to 900 ms. These 4 windows respectively reflect the processing of sound information during these time spans: 101 to 250 ms; 251 to 400 ms; 401 to 550 ms; 551 to 700 ms. The means and standard deviations of the POF data within these four windows are presented in Table III.

Repeated measures ANOVAs were carried out to examine if the POF (converted to R scores) on the four fixation locations (i.e., target, competitor, and distractors) were different across the four windows. The main effect of the fixation locations and the windows were significant [location: $F(3,69) = 124.07$, $p < 0.001$; window: $F(3, 69) = 30.27$, $p < 0.001$]. Specifically, to test the difference between POF on targets and on competitors in each window, paired sample t tests were also undertaken on the basis of subject (t_1) and item (t_2) variability.

Results from the t tests showed that for the *Low Tone 3* condition, the POFs on the targets (objects associated with Tone 3 words) were significantly higher than those on the competitors (objects associated with Tone 2 words) for all four windows [Window 1: Target mean = 0.426, competitor mean = 0.229, $t_1(23) = 4.6$, $p < 0.01$ and $t_2(7) = 3.38$, $p < 0.05$; Window 2: Target mean = 0.574, competitor mean = 0.178, $t_1(23) = 9.14$, $p < 0.01$ and $t_2(7) = 5.31$, $p < 0.01$; Window 3: Target mean = 0.665, competitor mean = 0.129, $t_1(23) = 11.49$, $p < 0.01$ and $t_2(7) = 5.98$, $p < 0.01$; Window 4: Target mean = 0.746, competitor mean = 0.101, $t_1(23) = 11.81$, $p < 0.01$ and $t_2(7) = 11.07$, $p < 0.001$].

In the *High Tone 3* condition, a preference for targets over competitors started to be significant from Window 2 [Window 2: Target mean = 0.463, competitor mean = 0.254, $t_1(23) = 5.2$, $p < 0.01$ and $t_2(7) = 3.82$, $p < 0.01$; Window 3: Target mean = 0.585, competitor mean = 0.216, $t_1(23) = 7.79$, $p < 0.01$ and $t_2(7) = 3.47$, $p = 0.01$; Window 4: Target mean = 0.690, competitor mean = 0.178, $t_1(23) = 10.86$, $p < 0.01$ and $t_2(7) = 4.71$, $p < 0.01$].

In the *Low Tone 2* condition, a significant fixation difference between competitors and targets only began showing in Window 2 [Window 2: Target mean = 0.246, competitor mean = 0.396, $t_1(23) = -2.34$, $p < 0.05$ and $t_2(7) = -1.77$, $p > 0.05$]. The two curves crossed over in Window 3. In

TABLE III. Experiment 1 means and standard deviations of POF data in four 150 ms windows.

		Window 1 (301 to 450 ms)	Window 2 (451 to 600 ms)	Window 3 (601 to 750 ms)	Window 4 (751 to 900 ms)
Low Tone 3 condition	Target (Tone 3)	0.426 (0.126)	0.574 (0.128)	0.665 (0.129)	0.746 (0.146)
	Competitor (Tone 2)	0.229 (0.106)	0.178 (0.098)	0.129 (0.087)	0.101 (0.083)
	Distractor 1	0.152 (0.068)	0.010 (0.070)	0.085 (0.077)	0.069 (0.069)
	Distractor 2	0.192 (0.095)	0.148 (0.086)	0.121 (0.073)	0.083 (0.055)
High Tone 3 condition	Target (Tone 3)	0.342 (0.101)	0.463 (0.131)	0.585 (0.144)	0.690 (0.128)
	Competitor (Tone 2)	0.279 (0.079)	0.254 (0.092)	0.216 (0.101)	0.178 (0.097)
	Distractor 1	0.168 (0.066)	0.130 (0.056)	0.085 (0.058)	0.055 (0.058)
	Distractor 2	0.211 (0.089)	0.154 (0.096)	0.113 (0.081)	0.077 (0.055)
Low Tone 2 condition	Target (Tone 2)	0.244 (0.130)	0.246 (0.143)	0.356 (0.152)	0.458 (0.210)
	Competitor (Tone 3)	0.318 (0.165)	0.396 (0.162)	0.317 (0.119)	0.245 (0.153)
	Distractor 1	0.184 (0.135)	0.173 (0.122)	0.173 (0.118)	0.143 (0.116)
	Distractor 2	0.254 (0.140)	0.186 (0.109)	0.154 (0.110)	0.154 (0.143)
High Tone 2 condition	Target (Tone 2)	0.247 (0.086)	0.257 (0.132)	0.343 (0.168)	0.478 (0.173)
	Competitor (Tone 3)	0.376 (0.128)	0.430 (0.127)	0.363 (0.154)	0.250 (0.159)
	Distractor 1	0.172 (0.107)	0.140 (0.090)	0.105 (0.090)	0.125 (0.086)
	Distractor 2	0.206 (0.109)	0.173 (0.139)	0.190 (0.122)	0.146 (0.091)

Window 4, judgments were securely locked on these later-decided targets, and the POF difference was once again significant, but in the opposite direction [Window 4: Target mean = 0.457, competitor mean = 0.245, $t_1(23) = 3.11$, $p < 0.01$ and $t_2(7) = 5.24$, $p < 0.01$].

A pattern similar to the *Low Tone 2* condition was observed in the *High Tone 2* condition [Window 1: Target mean = 0.247, competitor mean = 0.376, $t_1(23) = -3.79$, $p < 0.01$ and $t_2(7) = -2.05$, $p < 0.1$; Window 2: Target mean = 0.257, competitor mean = 0.430, $t_1(23) = -4.17$, $p < 0.001$ and $t_2(7) = -2.49$, $p < 0.05$]. The two curves crossed over in Window 3. The POF on targets was significantly higher compared to the POF on competitors in Window 4 [Window 4: Target mean = 0.478, competitor mean = 0.250, $t_1(23) = 3.73$, $p < 0.01$ and $t_2(7) = 4.26$, $p < 0.01$].

3. Analysis of diverging points

As these *t*-tests were carried out using data collapsed within 150 ms time windows, it is difficult to interpret the results in terms of how pitch information at specific time points influenced the POF data on targets and competitors. In order to reveal these effects in a fine-grained time scale, for each 2 ms time bin between 1 and 1200 ms from tone onset, 95% confidence intervals from each POF curve were obtained using a subjects-based bootstrap re-sampling procedure with 10 000 iterations (Efron and Tibshirani, 1993). The lower boundaries of the confidence intervals were then compared with chance (25%) to determine the time points at which POF on target or competitor was significantly above chance (see Fig. 4).

In the two low offset conditions, the lower boundaries of the confidence interval began to be above chance at 314 ms (in the *Low Tone 3* condition) and 346 ms (in the *High Tone 3* condition). In the two high offset conditions, the lower boundaries of the confidence interval of POF on competitor began to be above chance at 434 ms (in the *Low Tone 2* condition) and 348 ms (in the *High Tone 2* condition). This data consistently demonstrates a significant above chance preference to objects associated with Tone 3 words, observed approximately between 350 and 450 ms after tone onset. Considering the roughly 200 ms saccade latency before the fixations were observed, this suggests that the pitch information at 150 to 250 ms into the tone, which is around the turning point (200 ms into the tone), directed fixations to targets or competitors.

The lower boundary of POF on target (object associated with Tone 2 words) was above chance level at 668 ms (in the *Low Tone 2* condition) and 732 ms (in the *High Tone 2* condition). This result indicates the critical influence of the offset pitch on identification of the target, considering the approximately 500 ms tone duration in addition to the 200 ms saccade latency.

In summary, the results of Experiment 1 suggest:

(1) The critical point that influenced the on-line judgment of Tone 3 was the turning point, which has a pitch trough.
(2) In the two high offset conditions, the high offset pitch served as a pivotal signal for the final judgment of Tone 2.
(3) The final tone judgments were influenced by the one semitone pitch difference of tone offset.

III. EXPERIMENT 2

As noted above, Experiment 2 was similar to Experiment 1 with the exception that characters instead of pictures were used as the visual stimuli. We examined whether the response pattern would be the same regardless of this difference in the paradigm.

A. Method

1. Subjects

A different group of 22 native Mandarin speakers (13 males, 9 females) were recruited from international and visiting students and scholars at the University of California, San Diego. The average age of the subjects was 25.2 (standard deviation = 3.9). All were from Mainland China and had

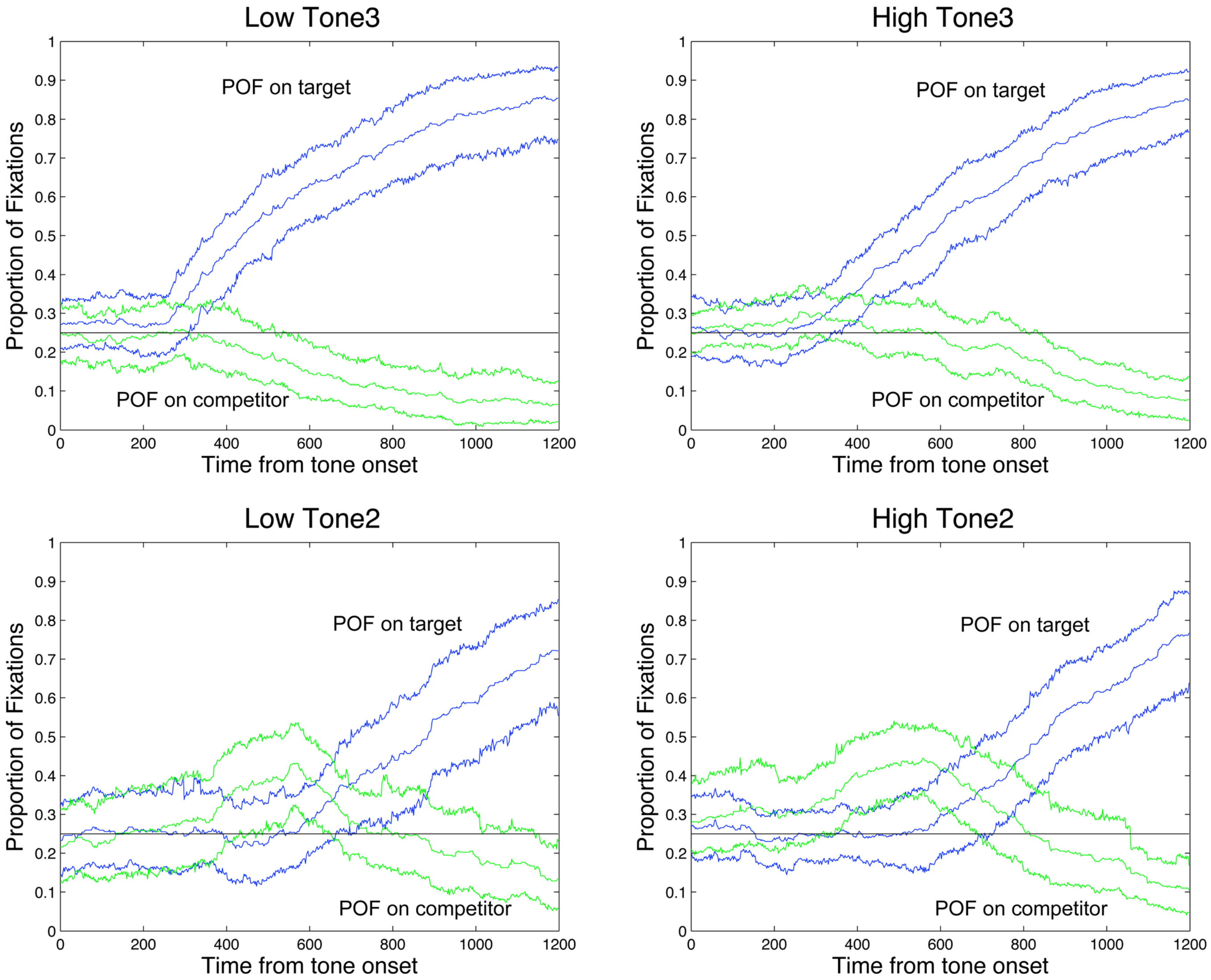

FIG. 4. (Color online) Experiment 1 bootstrapping curves in 2 ms interval (time scale in milliseconds). POF on target: the middle line is the mean target POF; the other two lines are the upper and lower boundaries of the 95% confidence interval of the target POF. POF on competitor: the middle line is the mean competitor POF; the other two lines are the upper and lower boundaries of the 95% confidence interval of competitor POF. Horizontal line is chance level (25%).

been living in the United States for a mean of 2.7 months and a range of 1 to 12 months. None of the subjects had more than 5 years of musical training, and none had any recent musical training within the past 10 years. None of them reported any hearing or speech disorders, and they all had normal or corrected to normal vision. None of them had participated in Experiment 1. Informed consent was obtained prior to the experiment, and the subjects were paid for their participation. Two additional subjects were tested but their data were excluded from all the analyses due to excessive eye blinking (over 30% of trials).

2. Stimuli and instrumentation

The sound stimuli were created in the same way as in Experiment 1, except that ten syllables were used to carry the Tone 2 and 3 tokens (see Table IV for their Pinyin and the characters). The consonants of these syllables were controlled in the same way as in Experiment 1. The visual stimuli were black Chinese characters on a gray background. They included ten Tone 3 characters, ten Tone 2 characters that shared the same syllable with the Tone 3 ones, ten Tone 1 characters, and ten Tone 4 characters. None of the Tone 1 and Tone 4 characters shared the same syllable as any of the Tone 2 and Tone 3 characters, and none of the Tone 1 characters shared the same syllable with any of the Tone 4 characters. They were all controlled for visual complexity (i.e., number of strokes) and lexical frequency. The majority of the characters (85%) did not correspond to any object used in Experiment 1. They were typed in Chinese Song font and resized to 120 × 120 pixels.

The instrumentation was identical to Experiment 1 except that an Eyelink II head-mounted tracker (SR Ltd., Canada) was used in this experiment.

3. Procedure

The experimental procedure was identical to that in Experiment 1, except that there was no name training block prior to the experiment.

TABLE IV. Sound stimuli used in Experiment 2 (visual stimuli: Chinese characters).

	Pinyin	Character
Tone 2 stimuli	/li/	离
	/lian/	联
	/mian/	棉
	/miao/	苗
	/wei/	围
	/wen/	闻
	/wu/	吴
	/yan/	言
	/yu/	鱼
	/yuan/	员
Tone 3 stimuli (same syllables as Tone 2 stimuli)	/li/	理
	/lian/	脸
	/mian/	免
	/miao/	秒
	/wei/	尾
	/wen/	稳
	/wu/	武
	/yan/	眼
	/yu/	雨
	/yuan/	远
Tone 1 (distractors)	/chao/	超
	/feng/	封
	/jing/	京
	/ke/	科
	/mo/	摸
	/pian/	偏
	/qian/	铅
	/tong/	通
	/yin/	音
	/zhuo/	桌
Tone 4 (distractors)	/dong/	洞
	/gu/	固
	/huo/	获
	/jia/	架
	/luo/	落
	/na/	纳
	/ruo/	弱
	/sheng/	盛
	/wang/	忘
	/xi/	细

B. Results and discussion

1. Tone identification data

Analysis of the tone identification data showed that the subjects had an overall rate of correct response of 90.1% at the end of the syllables. The breakdown of the correct response rate for the 4 conditions were 95.5% (selecting Tone 3 in the *Low Tone 3* condition); 91.9% (selecting Tone 3 in the *High Tone 3* condition); 83.9% (selecting Tone 2 in the *Low Tone 2* condition); and 90.6% (selecting Tone 2 in the *High Tone 2* condition). For further analyses, the selected characters are termed targets; the characters associated with the same syllables as the targets but different tones are termed competitors; the Tone 1 and 4 characters that were presented on the same screen are termed distractors.

Similar to the results of Experiment 1, Tone 3 characters were more frequently selected in the low offset conditions, and Tone 2 characters in the high offset conditions [$ps < 0.01$]. Tone 3 was also more frequently selected in the *Low Tone 3* condition compared to the *High Tone 3* condition [$t(21) = 2.26$, $p < 0.05$], and Tone 2 was more often selected in the *High Tone 2* condition compared with the *Low Tone 2* condition [$t(21) = 5.09$, $p < 0.01$].

2. Temporal window analysis

The eye fixation data were processed in the same way as in Experiment 1. Those trials in which the subjects selected characters other than the targets (9.9%) were excluded from further analysis of the eye fixation data. The trials that contained blinks (7.7%) during presentation of the sound tokens were also excluded from analysis of the eye fixation data. The POF data was collapsed over subjects and items, to create averaged trajectory graphs in 20 ms time windows (see Fig. 5), for each of the 4 conditions.

By visual inspection, the POF curves in this experiment appear to have a pattern similar to that of Experiment 1, so the same 4 windows were adopted for performing ANOVAs (the time intervals of these 4 windows are: 301 to 450 ms; 451 to 600 ms; 601 to 750 ms; 751 to 900 ms). The means and standard deviations of the POF data in these four windows for the four conditions are presented in Table V.

The same repeated measures ANOVAs on the transformed POF data showed that the POF on the four fixation locations (i.e., target, competitor, and two distractors) were different across the four windows [location: $F(3,63) = 79.81$, $p < 0.001$; window: $F(3, 63) = 37.93$, $p < 0.001$]. Paired sample t tests that were identical to those in Experiment 1 were carried out on the basis of subject (t_1) and item (t_2) variability.

The results of t tests showed that for the *Low Tone 3* condition, POF on the targets (Tone 3 words) were significantly higher than those on the competitors (Tone 2 words) beginning on Window 2 [Target mean = 0.399, competitor mean = 0.250, $t_1(21) = 3.80$, $p < 0.01$ and $t_2(9) = 4.40$, $p < 0.01$] and continued significant thereafter [Window 3: Target mean = 0.499, competitor mean = 0.248, $t_1(21) = 5.13$, $p < 0.001$ and $t_2(9) = 3.67$, $p < 0.01$; Window 4, Target mean = 0.587, competitor mean = 0.210, $t_1(21) = 6.31$, $p < 0.001$ and $t_2(9) = 5.92$, $p < 0.001$].

The pattern in the *High Tone 3* condition was similar to that in the *Low Tone 3* condition [Window 2: Target mean = 0.359, competitor mean = 0.271, $t_1(21) = 2.80$, $p < 0.05$ and $t_2(9) = 1.32$, $p > 0.05$; Window 3: Target mean = 0.430, competitor mean = 0.249, $t_1(21) = 3.81$, $p < 0.01$ and $t_2(9) = 2.41$, $p < 0.05$; Window 4: Target mean = 0.502, competitor mean = 0.199, $t_1(21) = 6.00$, $p < 0.001$ and $t_2(9) = 4.71$, $p < 0.01$].

In the *Low Tone 2* condition, the preference for competitors (Tone 3 words) over the targets (Tone 2 words) began to show in Window 2 [target mean = 0.280, competitor mean = 0.343, $t_1(21) = -2.21$, $p < 0.05$; $t_2(9) = -1.95$, $p = 0.08$]. The two curves then crossed over in Window 3 and significantly more fixations were then on targets than on competitors

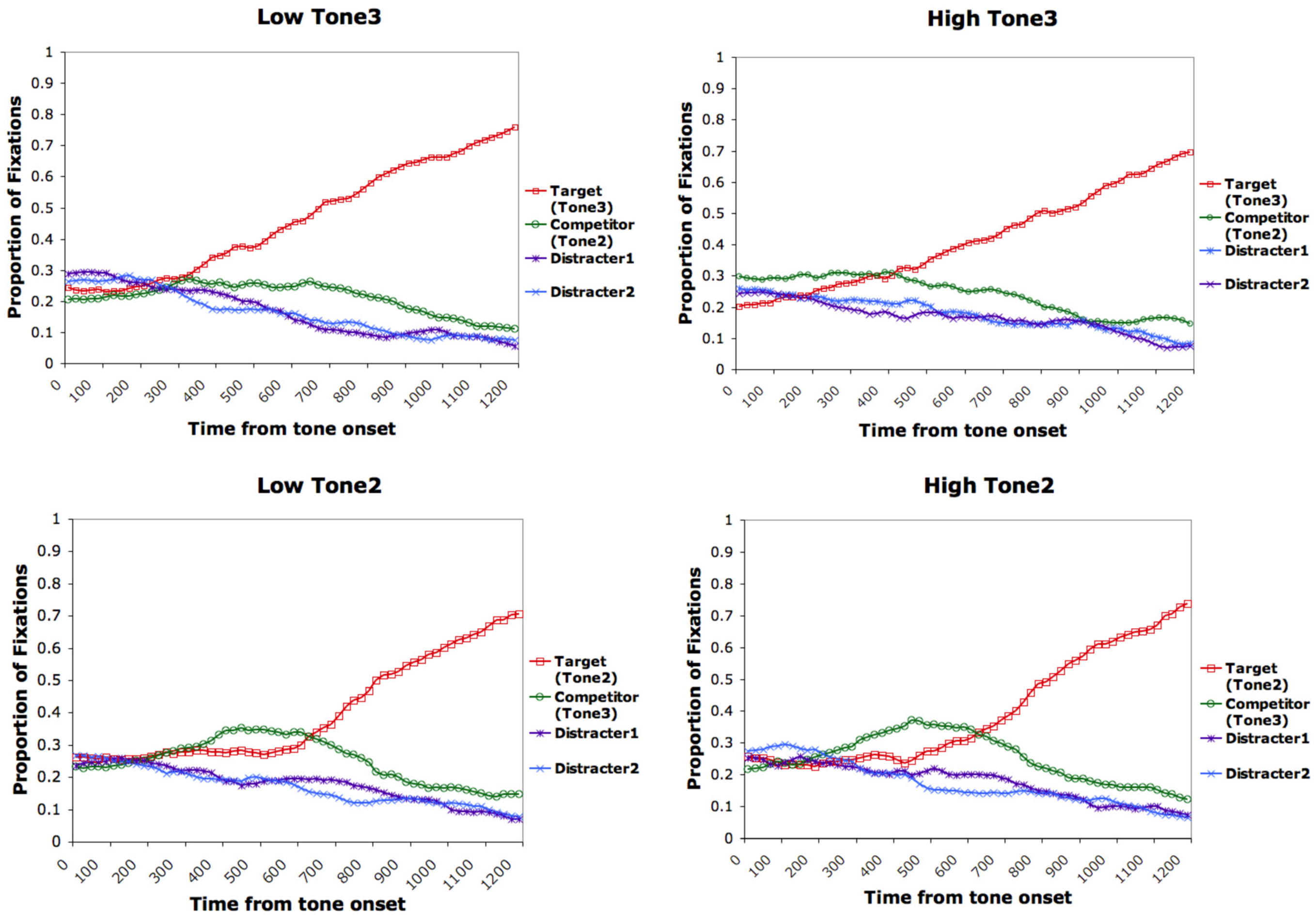

FIG. 5. (Color online) Experiment 2 POF curves in 20 ms interval (time scale in milliseconds).

[Window 4: Target mean = 0.499, competitor mean = 0.221, $t_1(21) = 5.93$, $p < 0.001$ and $t_2(9) = 7.15$, $p < 0.001$].

The pattern in the *High Tone 2* condition was again similar to the one in the *Low Tone 2* condition [Window 1: Target mean = 0.252, competitor mean = 0.334, $t_1(21) = -2.80$, $p < 0.05$ and $t_2(9) = -1.28$, $p > 0.05$; Window 2: Target mean = 0.284, competitor mean = 0.356, $t_1(21) = -2.58$, $p < 0.05$ and $t_2(9) = -1.79$, $p > 0.05$]. The two curves crossed over in Window 3. The POF on targets was significantly higher compared to the POF on competitors in Window 4 [Window 4: Target mean = 0.505, competitor mean = 0.214, $t_1(21) = 6.25$, $p < 0.001$ and $t_2(9) = 3.97$, $p < 0.01$].

TABLE V. Experiment 2 means and standard deviations of POF data in four 150 ms windows.

		Window 1 (301 to 450 ms)	Window 2 (451 to 600 ms)	Window 3 (601 to 750 ms)	Window 4 (751 to 900 ms)
Low Tone 3 condition	Target (Tone 3)	0.321 (0.119)	0.398 (0.122)	0.495 (0.139)	0.588 (0.153)
	Competitor (Tone 2)	0.260 (0.088)	0.250 (0.083)	0.248 (0.094)	0.210 (0.109)
	Distractor 1	0.229 (0.087)	0.082 (0.066)	0.118 (0.059)	0.091 (0.046)
	Distractor 2	0.190 (0.078)	0.169 (0.073)	0.139 (0.082)	0.111 (0.078)
High Tone 3 condition	Target (Tone 3)	0.300 (0.089)	0.359 (0.102)	0.430 (0.127)	0.502 (0.157)
	Competitor (Tone 2)	0.304 (0.084)	0.271 (0.070)	0.249 (0.108)	0.199 (0.096)
	Distractor 1	0.217 (0.102)	0.197 (0.108)	0.158 (0.083)	0.146 (0.097)
	Distractor 2	0.180 (0.077)	0.174 (0.086)	0.164 (0.080)	0.153 (0.093)
Low Tone 2 condition	Target (Tone 2)	0.280 (0.074)	0.280 (0.086)	0.361(0.133)	0.499 (0.151)
	Competitor (Tone 3)	0.320 (0.108)	0.343 (0.080)	0.305 (0.081)	0.221 (0.091)
	Distractor 1	0.205 (0.070)	0.187 (0.067)	0.189 (0.078)	0.154 (0.088)
	Distractor 2	0.195 (0.085)	0.190 (0.198)	0.144 (0.072)	0.127(0.080)
High Tone 2 condition	Target (Tone 2)	0.252 (0.077)	0.284 (0.085)	0.361 (0.113)	0.505 (0.156)
	Competitor (Tone 3)	0.334 (0.097)	0.356 (0.074)	0.305 (0.072)	0.214 (0.076)
	Distractor 1	0.209 (0.086)	0.205 (0.080)	0.190 (0.077)	0.144 (0.074)
	Distractor 2	0.205 (0.075)	0.155 (0.069)	0.143 (0.079)	0.137 (0.093)

3. *Analysis of diverging points*

The same bootstrapping procedure as in Experiment 1 was used for Experiment 2 (see Fig. 6 for the curves).

In the *Low Tone 3* and *High Tone 3* conditions, the POF curves that represent fixations on Tone 3 characters began rising by 300 ms into the tone. The lower boundaries of the confidence interval began to be above chance from 382 and 416 ms, respectively, for these two conditions, which indicates that the sound signal around the turning point had a critical influence on the preference toward the target.

In the *Low Tone 2* and *High Tone 2* conditions, the lower boundary of the confidence interval was above chance following 388 ms in the *Low Tone 2* condition, and 348 ms in the *High Tone 2* condition from tone onset. Approximately after 500 ms into the tone, the fixations on the competitor began to decrease and the ones on target to increase. The lower boundary of the confidence interval on target (Tone 2 words) was above chance starting from 656 ms (*Low Tone 2* condition) and 624 ms (*High Tone 2* condition) into the tone, which aligned with the time point of approximately 200 ms following tone offset.

IV. GENERAL DISCUSSION

Using the visual world paradigm, the present findings demonstrated the dynamic process of lexical tone perception by revealing the effect of fine-grained pitch information at tone offset and turning point on on-line tone judgment for native Mandarin-speaking listeners. The eye fixation data in both experiments yielded a very similar pattern. Low turning point pitch was exploited as a cue for on-line identification of Tone 3; offset pitch height was a critical cue for disambiguating Tones 2 and 3.

The temporal window and diverging point analyses both showed that the difference between the POFs on Tone 3 versus Tone 2 items did not become significant until around 400 ms; this revealed the pivotal influence of the low pitch trough of the turning point on Tone 3 perception. Prior studies on Mandarin tone perception and production have

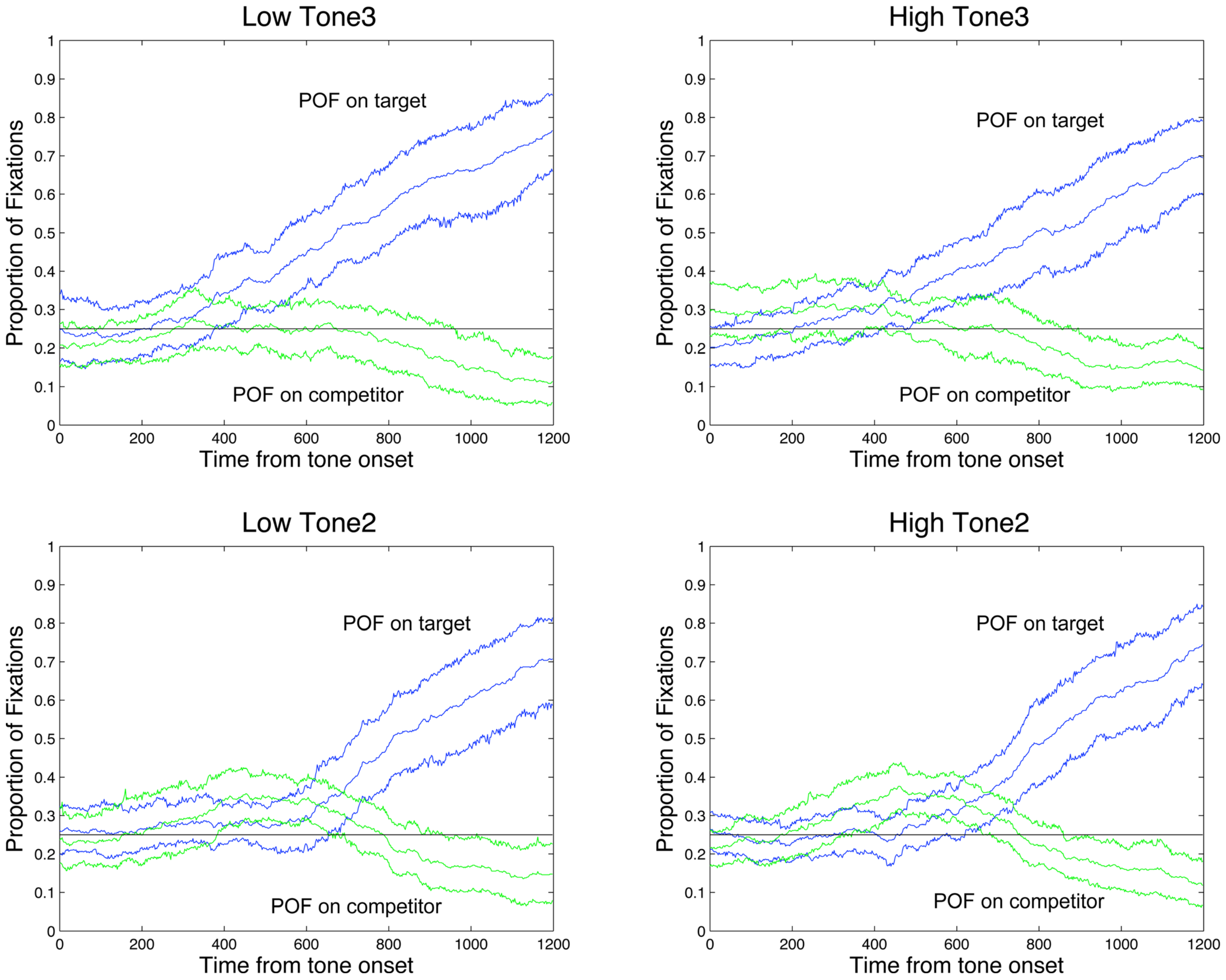

FIG. 6. (Color online) Experiment 2 bootstrapping curves in 2 ms interval (time scale in milliseconds). POF on target: the middle line is the mean target POF; the other two lines are the upper and lower boundaries of the 95% confidence interval of target POF. POF on competitor: the middle line is the mean competitor POF; the other two lines are the upper and lower boundaries of the 95% confidence interval of competitor POF. Horizontal line is chance level (25%).

suggested that this low pitch target is a critical parameter of Tone 3 (Xu and Wang, 2001; Wong *et al.*, 2005). Taking a broad view, the importance of turning point pitch in Mandarin Tone 3 is also in line with data from another tone language, Thai (Zsiga and Nitisaroj, 2007), in suggesting that pitch inflection at the syllable midpoint could provide perceptual cues for tone identification.

In the present study, pitch information at tone onset (1 to 100 ms) initially directed more fixations to Tone 3 words but did not provide perceptual evidence strong enough to influence tone judgment significantly. At first sight, this finding appears inconsistent with the results of Gottfried and Suiter (1997) and Lee *et al.* (2008) since these authors used only the initial 6 pitch cycles (about 30 ms in duration) of the syllable, and native speakers were able to correctly identify the tones in most cases. However, when examined more closely, the data from both the above studies also showed that if given only the tone onset (without offset), native speakers tend to confuse Tone 2 with Tone 3, because these two tones share the characteristic of low onset pitch. Since the present study focused on word pairs consisting of Tones 2 and 3, and their onset pitches were very close to each other (around 185 Hz for Tone 3, 200 Hz for Tone 2), our results are consistent with these two studies in showing that native speakers had difficulty in identifying the sound token as a Tone 3 word based only on the cue of low onset pitch. These findings also align with the literature showing that Tones 2 and 3 are particularly difficult to differentiate, even for native speakers (Blicher *et al.*, 1990; Shen *et al.*, 1993; Whalen and Xu, 1992).

To tease apart the influence of the rising slope and the offset pitch height on on-line tone judgment, we examined the change in fixations on Tone 2 items across time. Because the steep slope of the rising pitch had fully unfolded by 400 ms into the tone, and on average it took 200 ms or less to program and execute a saccade, the influence of the rising slope should appear by 600 ms into the tone. This was indeed shown in both experiments by POF curves that represent fixations on Tone 2 items, which began rising around 500 to 600 ms. This result is in accordance with the earlier finding that a steep slope of pitch change is utilized as a cue for identifying Tone 2 (Xu *et al.*, 2006). However, the diverging point analysis showed that a preference for target Tone 2 became significant only when information concerning offset pitch was available, which suggests that the cue of slope works better for tone identification when it is integrated with endpoint pitch height information.

The ultimate tone judgment data replicates findings (Gottfried and Suiter, 1997; Lee *et al.*, 2008) showing that native speakers of Mandarin identify tokens with low onset and high offset pitch mostly as Tone 2 and those with low onset and low offset pitch as Tone 3. Overall, the mean correct rate was higher for the two low offset conditions (96.4%) compared with the two high offset conditions (79%). This is likely due to the ambiguity in the two high offset conditions introduced by the hybrid pattern consisting of the Tone 3 onset mixed with Tone 2 offset. It was also found the listeners did respond to the difference between the original and the ambiguous conditions, which only had one semitone difference at the offset. Those conditions with original offset pitch height (i.e., *Low Tone 3* and *High Tone 2* conditions) were identified as the corresponding tones in 90% of the trials. This occurred significantly more often compared to the ambiguous conditions (i.e., *High Tone 3* and *Low Tone 2* conditions). Overall the sound tokens in the ambiguous conditions were identified as the corresponding tones in 85% of the trials. This finding is consistent with Shen *et al.* (2011) in showing native speakers' high sensitivity to small pitch differences in tone perception.

As reviewed by Dahan and Magnuson (2006), spoken word recognition is an incremental and constantly changing process, in which listeners continually evaluate the unfolding of the speech stream with fine-grained sensitivity, and activate certain lexical candidates that compete for recognition, even from the initiation of an utterance. The present study is in accordance with this line of research by showing that even short segments of pitch information can influence on-line tone judgment before the syllable ends. Because the listeners had a time interval of 700 ms to rapidly skim through the visually displayed objects or characters at the beginning of a trial during presentation of the prompt sentence, they already knew the location of the corresponding visual stimulus when the sound token was played, and so were able to direct attention (and make a saccade) instantly to this item. In both experiments, sound stimuli were manipulated in such a way that the meaning of the spoken word depended on pitch cues at critical points in the syllable (e.g., a low turning point pitch indicates Tone 3 instead of Tone 2). In order to follow the instruction, the listeners had to exploit these pitch cues, make instantaneous saccadic eye movements to those lexical candidates before clicking on them. Aggregated over a large amount of trials, the listeners' eye fixation data (i.e., which object or character was fixated on at a certain time point) revealed influences of these fine-grained pitch cues on tone judgments.

By using different types of visual stimuli (object pictures versus Chinese characters) in the visual world paradigm, the two experiments reported here demonstrate a similar time trajectory of eye fixations on the tone targets and competitors (see Figs. 4 and 6). It suggests that when orthographic information is controlled (e.g., number of strokes that represents visual complexity), the written forms of words in a logographic language may be used as visual stimuli in a visual world study that investigates lexical tones. However because the two experiments were carried out using two different sets of items and two different groups of subjects, it is difficult to make a direct comparison of the results from using the two types of visual stimuli. This issue will be investigated in future research.

V. CONCLUSION

In summary, using an eye-tracking paradigm for examining time-sensitive perceptual processing, the present study demonstrates the importance of pitch height at tone offset and turning point in the process of tone identification by native speakers of Mandarin. The study also extends research on spoken word recognition by providing evidence that fine-grained pitch information influences the dynamic process of tone perception as the syllable unfolds, even before the tone judgment is ultimately made.

ACKNOWLEDGMENTS

We thank Jinmian Yang, Mackenzie Glaholt, Liz Schotter, Bernhard Angele, Tim Slattery, David Huber, and Don MacLeod for their assistance and helpful comments on the paper. The research was partially supported by an Atkinson Endowment.

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.

Boersma, P. and Weenink, D. (**2008**). "Praat: Doing phonetics by computer (Version 5.0.42)" [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/.

Chan, S. W., Chuang, C.-K., and Wang, W. S-Y. (**1975**). "Cross-linguistic study of categorical perception for lexical tone," J. Acoust. Soc. Am. **58**, S119.

Chandrasekaran, B., Gandour, J. T., and Krishnan, A. (**2007**). "Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity," Restor. Neurol. Neuros. **25**, 195–210.

Dahan, D., and Magnuson, J. S. (**2006**). "Spoken word recognition," in *Handbook of Psycholinguistics*, 2nd ed., edited by M. J. Traxler and M. A. Gernsbacher (Academic Press, Amsterdam), pp. 249–284.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (**2001**). "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition," Lang. Cognit. Processes **16**, 507–534.

Deutsch, D. (**1992**). "Some new pitch paradoxes and their implications," Philos. Trans. R. Soc. London, Ser. B **336**, 391–397.

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., Le, J., Shen, J., and Henthorn, T. (**2009**). "The pitch levels of female speech in two Chinese villages," J. Acoust. Soc. Am. **125**, 208–213.

Dolson, M. (**1994**). "The pitch of speech as a function of linguistic community," Music Percept. **11**, 321–331.

Efron, B., and Tibshirani, R. (**1993**). *An Introduction to the Bootstrap* (Chapman and Hall, New York), pp. 168–177.

Fox, R., and Qi, Y. (**1990**). "Contextual effects in the perception of lexical tone," J. Chin. Linguist. **18**, 261–283.

Francis, A. L., Ciocca, V., and Ng, B. K.-C. (**2003**). "On the (non)categorical perception of lexical tones," Percept. Psychophys. **65**, 1029–1044.

Francis, A., Ciocca, V., Wong, N., Leung, W., and Chu, P. (**2006**). "Extrinsic context affects perceptual normalization of lexical tone," J. Acoust. Soc. Am. **119**, 1712–1726.

Fu, Q. J., and Zeng, F. G. (**2000**). "Identification of temporal envelope cues in Chinese tone recognition," Asia Pac. J. Speech Lang. Hear. **5**, 45–57.

Gandour, J. (**1983**). "Tone perception in Far Eastern languages," J. Phonetics **11**, 149–175.

Gottfried, T. L., and Suiter, T. L. (**1997**). "Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones," J. Phonetics. **25**, 207–231.

Halle, P. A., Chang, Y.-C., and Best, C. T. (**2004**). "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," J. Phonetics **32**, 395–421.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**, 3983–3994.

Huetting, F., and Altmann, G. T. M. (**2007**). "Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness," Vis. Cogn. **15**, 985–1018.

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

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. (**1985**). "Tone perception," J. Chin. Linguist. **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.

Malins, J., and Joanisse, M. F. (**2010**). "The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study," J. Mem. Lang. **62**, 407–420.

Massaro, D. W., Cohen, M. M., and Tseng, C. (**1985**). "The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese," J. Chin. Linguist. **13**, 267–290.

Matin, E. (**1974**). "Saccadic suppression: A review and an analysis," Psychol. Bull. **81**, 899–917.

McCawley, J. C. (**1978**). "What is a tone language?" in *Tone: A Linguistic Survey*, edited by V. Fromkin (Academic Press, New York), pp. 113–131.

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (**2002**). "Gradient effects of within-category phonetic variation on lexical access," Cognition **86**, B33–B42.

Moore, C., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.

Rayner, K. (**1998**). "Eye movements in reading and information processing: 20 years of research," Psychol. Bull. **124**, 372–422.

Rayner, K. (**2009**). "Eye movements and attention in reading, scene perception, and visual search," Q. J. Exp. Psychol. **62**, 1457–1506.

Rayner, K., and Clifton, C. (**2009**). "Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research," Biol. Psychol. **80**, 4–9.

Shen, J., Deutsch, D., and Le, J. (**2011**). "Overall pitch height as a cue for lexical tone perception," poster session presented at the 162nd Meeting of Acoustical Society of America, San Diego, CA.

Shen, X., Lin, M., and Yan, J. (**1993**). "*F*0 turning point as an *F*0 cue to tonal contrast: A case study of Mandarin tones 2 and 3," J. Acoust. Soc. Am. **93**, 2241–2243.

Speer, S. R., and Xu, L. (**2005**). "Ambiguous lexical tone process during Mandarin sentence comprehension-evidence form eye-movement experiment," paper presented at the *11th Annual Conference on Architectures and Mechanisms for Language Processing*, Ghent, Belgium.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res. **28**, 455–462.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., and Chambers, C. (**2000**). "Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing," J. Psycholinguist. Res. **29**, 557–580.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (**1995**). "Integration of visual and linguistic information in spoken language comprehension," Science **268**, 1632–1634.

Wang, W. S.-Y. (**1973**). "The Chinese language," Sci. Am. **228**, 50–60.

Wang, W. S.-Y. (**1976**). "Language change," Ann. N.Y. Acad. Sci. **280**, 61–72 (1976).

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., Schwartz, R. G., and Jenkins, J. J. (**2005**). "Perception and production of lexical tones by 3-year-old, Mandarin-speaking children," J. Speech. Lang. Hear. Res. **48**, 1065–1079.

Wong, P. C.-M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech Lang. Hear. Res. **46**, 413–421.

Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**, 1063–1074.

Xu, Y., and Sun, X. (**2001**). "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. **111**, 1399–1413.

Xu, Y., and Wang, E. (**2001**). "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**, 319–337.

Zhou, X., Shu, H., Bi, Y., and Shi, D. (**1999**). "Is there phonologically mediated access to lexical semantics in reading Chinese?" in *Reading Chinese Script: A Cognitive Analysis*, edited by J. Wang, A. Inhoff, and H.-C. Chen (Erlbaum, Hillsdale, NJ), pp. 135–171.

Zsiga, E., and Nitisaroj, R. (**2007**). "Tone features, tone perception, and peak alignment in Thai," Lang. Speech **50**, 343–383.