J Sign Process Syst (2016) 82:253–261
DOI 10.1007/s11265-015-1008-2

Context Effect in the Categorical Perception of Mandarin Tones

Fei Chen[1] · **Gang Peng**[1,2]

Received: 14 November 2014 / Revised: 6 March 2015 / Accepted: 21 April 2015 / Published online: 6 May 2015

Abstract The categorical perception of tones is based not only on word-internal F0 cues but also on external F0 cues in the contexts. The present study focuses on the effects of different types of preceding contexts on Mandarin tone perception. In the experiment, subjects were required to identify a target tone with the preceding context. The target tone was from a tone continuum ranging from Mandarin Tone 1 (high-level tone) to Tone 2 (mid-rising tone). It was preceded by four types of contexts (normal speech, reversal speech, fine-structure sound, and non-speech) with different mean F0 values. Results indicate that the categorical perception of Mandarin tones is influenced only by the normal speech context, and the effect is contrastive. For instance, in a normal speech context with a higher mean F0, the following tone is more likely to be perceived as a lower-frequency tone (Tone 2), whereas with a lower mean F0, the following tone is more likely to be perceived as a higher-frequency tone (Tone 1). These findings suggest that Mandarin tone normalization is mediated by speech-specific processes and that the speech context needs to be intelligible.

Keywords Mandarin tone · Context effect · Categorical perception · Speech-specific mechanisms

✉ Gang Peng
gpengjack@gmail.com

Fei Chen
chenfei@siat.ac.cn

[1] Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2] Department of Linguistics and Modern Languages, and Joint Research Centre for Language & Human Complexity, The Chinese University of Hong Kong, Hong Kong SAR, China

1 Introduction

Tone languages such as Mandarin use pitch patterns to distinguish lexical meanings [1], and fundamental frequency (F0) is the most important physical correlate of pitch. As can be seen in Fig. 1, Mandarin Chinese has four different lexical tones: high-level tone (Tone 1), mid-rising tone (Tone 2), low-falling-then-rising tone (Tone 3), and high-falling tone (Tone 4). The same syllable "ma" with different lexical tones will have very different meanings, e.g., "mother" (Tone 1), "hemp" (Tone 2), "horse" (Tone 3), and "scold" (Tone 4) respectively.

However, it is worth noting that in actual speech the exact F0 values of lexical tones are highly variable across utterances and across talkers. Abundant literature has shown that there is a great deal of inter- and intra-talker variability in speech production, giving rise to varied F0 realizations of tone [2]. For instance, the same word uttered by different talkers (i.e., inter-talker variability) or by the same talker on different occasions (i.e., intra-talker variability) may differ significantly in terms of their acoustic properties. How then do listeners deal with such inter- and intra-talker differences in F0?

The term "tone normalization" has been used to describe the processes by which listeners recognize the same tone produced by different talkers or the same talker in different conditions [3]. There are two types of cues mainly used in these processes – word-internal cues and word-external cues. Wang [1] discussed various word-internal cues that might be available during tone normalization, including F0, duration, intensity profile, voice quality, and other relevant acoustic cues of a word. All of these acoustic cues contain useful information about the tone category, among which F0 is the most important cue for tone perception. As for word-external cues, they mainly refer to acoustic cues in the context, and are thus also known as "contextual cues". Often, listeners make use of both

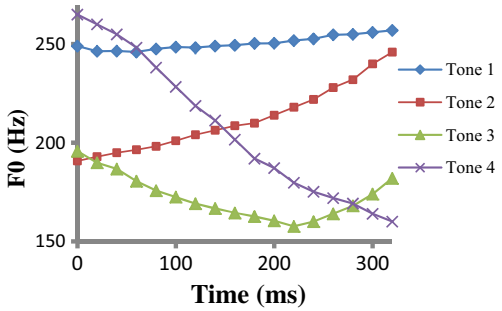


Figure 1 F0 contours of four Mandarin tones on the isolated syllable /ma/, produced by a talker of the present experiment.

word-internal F0 cues and contextual F0 cues for tone normalization [3–6]. But on some occasions, contextual cues with information about a talker's F0 range can be more crucial, such as when the stimuli to be categorized are highly ambiguous because they are located close to the perceptual boundaries of two tones.

1.1 The Effect of Speech Context on Lexical Tone Perception

The speech context with cues of a talker's F0 exerts an important influence on the perception of target lexical tones, which can be divided further into two categories: level tones that vary in F0 height but have similar contours, or contour tones which can be differentiated in terms of both F0 height and F0 direction.

Comparatively speaking, the effects of the speech context are most evident on level tones. Studies [3–5] have clearly demonstrated that the perception of Cantonese level tones is context-dependent. Francis et al. [3] found that target stimuli were also more likely to be perceived as low-level tones when positioned in a synthesized context with high F0, whereas the same set of stimuli were perceived as high-level tones in a synthesized context with low F0. Moreover, Listeners' tonal judgments were proportional to the degree of frequency shift. Wong and Diehl [4] used the three level tones from Cantonese (Tone 1: high-level tone, Tone 3: mid-level tone, and Tone 6: low-level tone) as target stimuli. They asked listeners to judge the identity of these tones with speech contexts which were manipulated to differ in mean F0 height. They found that the same target stimuli were identified as Tone 1 (high-level) 99.5 % of the time when in a lowered F0 context, Tone 6 (low-level) 95.8 % of the time when in a raised F0 context, and Tone 3 (mid-level) 91.9 % of the time when the context had an intermediate mean F0. Zhang et al. [5] further demonstrated that raised or lowered speech context conditions could lead to similar contrastive effects, regardless of the F0 contour being preserved or flattened.

In contrast to level tone perception, however, much more mixed results have been found for the influence of speech context on contour tone perception. Some results showed no significant context effect on Mandarin tone perception. Leather [7] tested the perception of syllables produced with Mandarin Tone 1 and Tone 2 following natural spoken sentences. Little information about the direction of perceptual shift was provided in his study, making it difficult to judge the directionality of any potential influence of context on lexical tone perception. Fox and Qi [8] examined the perception of Mandarin syllables that varied in F0 onset frequency from Tone 1 to Tone 2, both in isolation and when paired with a preceding syllable that had a typical Tone 1 or Tone 2 with a fixed F0. Native Mandarin listeners were asked to identify the tone category of the second syllable. There was only a small and inconsistent difference between perceptions of syllables in isolation and in context, and the effect was assimilatory. However, most studies revealed a contrastive context effect. In Lin and Wang's [9] study, the F0 of the first syllable was held constant with a typical F0 value for Tone 1 (high-level), while the onset F0 of the second syllable was manipulated with difference frequencies. Native Mandarin Chinese participants were asked to identify the tone category of the first syllable. Results showed that as the onset F0 of the second syllable increased, participants were more likely to label the first syllable as Tone 2 (Tone 2 has relatively lower F0 in comparison to Tone 1). The effect was contrastive. In another study, Moore and Jongman [10] investigated talker normalization in the perception of Mandarin Tone 2 (mid-rising) and Tone 3 (low-falling–rising), showing identification shift such that identical stimuli were identified as the low tone (Tone 3) for the high F0 precursor condition, but as high tone (Tone 2) for the low F0 precursor condition, also in a contrastive manner. Moreover, the latest research examined the effect of speech context of natural meaningful sentences on Mandarin tone perception. Huang and Holt [6] embedded different tone stimuli ranging over a continuum of eight steps, starting at Mandarin Tone 1 and ending at Tone 2, with speech sentence contexts produced by the same talker. With a speech context of higher mean F0, the following syllables were more likely to be perceived as a tone with lower-frequency (Tone 2), and the reversed pattern for syllables preceded by lower-F0 contexts.

In conclusion, the effects of speech context are much more evident and consistent on level tones than on contour tones, and consistently appear to be in a contrastive manner. However, the effects of speech context on contour tones are mixed and inconsistent, although in most cases they are also in a contrastive manner.

1.2 The Effect of Different Types of Contexts

A wide range of studies on phonetic speech categorization (vowel and consonant perception) have shown that higher frequency contexts shift perception towards a lower frequency phonetic alternative and vice versa [11–14]. This is highly

comparable to the contrastive effect of speech contexts on lexical tone normalization.

However, it is still unclear what effects non-speech contexts have on lexical tone normalization. Francis et al. [3] found that the F0 trajectory, extracted from the speech context and superimposed on a [ ] sound generated with the "hummed" neutral vocal tract, had no effect on the perception of Cantonese level tones, unlike the corresponding speech context. Listeners seemed unable or unwilling to use a linguistically meaningless context for the purposes of tone normalization. Despite some uncertainty about the interpretation of the [ ] sound (i.e., whether it is a non-speech or non-native speech sound for Cantonese listeners), this study casts doubt on the general perceptual mechanisms for tone normalization. Moreover, Zhang et al. [5] found that non-speech contexts constructed by triangle wave sounds (which have a different harmonic structure than normal speech sounds) only mildly changed the identification preference, revealing unequal effects of speech and non-speech contexts on the perceptual normalization of Cantonese level tones. Nevertheless, Huang and Holt [6, 15] placed an eight-step tone continuum (ranging from Mandarin Tone 1 to Tone 2) after non-speech contexts (either a harmonic tone or a pure tone context), and reported that the contrastive effect was also elicited by the non-speech precursors (which modeled the mean spectrum of F0 of speech) in a similar manner to the speech context.

How can the contradictory findings in previous studies be explained? One obvious difference is in the type of target lexical tones that were examined (Mandarin vs. Cantonese; contour tones vs. level tones). As mentioned above, effects among level tones are much more evident in a speech context in comparison to a non-speech context [3, 5]. However, speech and non-speech contexts exhibited qualitatively similar results for the contour tones [6, 15]. This discrepancy casts doubt on the previous results of Huang and Holt [6, 15] regarding the effect of preceding non-speech context on Mandarin lexical tones.

Obviously, tone differences are not the only factor contributing to these contradictory findings. The methodology for measuring the effects of contexts on the perception of target tones is quite different as well. Huang and Holt [6] used the paradigm of "categorical perception," in which an eight-step tone continuum was attached to the non-speech contexts. The effect of non-speech context was found to be greatest for the most ambiguous tone stimuli between Tone 1 and Tone 2 (i.e., stimuli in the middle of the continuum), where the higher mean F0 context mildly increased the percentage of Tone 2 responses. There was, however, barely any effect for the unambiguous tone stimuli (i.e., stimuli close to the two ends of the continuum). By contrast, Francis et al. [3] used a rather different method. The target tone stimulus was a mid-level tone, and the participants' responses were recorded and scored with a "count-sum" scoring system: If a subject gave a low level tone response, then the response was given a score of −1; mid-level tone responses were scored as 0; and high level tone responses were given a score of +1. Zhang et al. [5] used a similar method: following the perceptual scale of Cantonese level tones, a high-level tone response was coded as "6", a middle-level tone response as "3", and a low-level tone response as "1", with 6, 3, and 1 referring to level tones' perceptual heights, respectively. This analysis has the advantage of precisely estimating the overall change of tone responses according to F0 shift in a context condition.

There is also a big difference in the so-called non-speech materials. Huang and Holt [6, 15] made use of harmonics, specifically a sequence of 17 tone complexes composed of four equal-amplitude sine-waves with frequencies at the first four multiples of the F0. By contrast, Zhang et al. [5] use a triangle wave as their non-speech material, which has a different harmonic structure than do speech sounds. Finally, the [ ] sound used by Francis et al. [3] cannot actually be regarded as a non-speech sound, even though it was unintelligible to the Cantonese listeners. Hence, to make a clearer comparison between speech and non-speech materials used in the above studies, the contexts in the present study include both typical non-speech sounds as well as other linguistically unintelligible sound materials.

1.3 Research Aims

The overall purpose of this study is to investigate how different types of extrinsic contexts affect the perceptual normalization of Mandarin tones. More specifically, two questions are asked as follows:

First, given the mixed and inconsistent results regarding the effects of extrinsic contexts on Mandarin lexical tone perception, the present study aims to obtain a more reliable evaluation of the effects of the precursor contexts on Mandarin tone perception. We adopted Mandarin Tone 1 (high level) and Tone 2 (mid-rising) as target tones and investigated whether the effects of extrinsic contexts on Mandarin tones are contrastive or assimilatory.

Second, the present study investigates the potential mechanisms of lexical tone normalization. It is currently unclear whether the processing of supra-segmental properties of speech (such as the lexical tones) recruits speech-specific mechanisms, as suggested by some studies [3, 5], or general perceptual mechanisms, as proposed by others [6, 15].

In order to answer the first question, we adopted the paradigm of "categorical perception" as also used by Huang and Holt [6]. By manipulating the onset F0, eleven intermediate target tone stimuli were created on a continuum ranging from Mandarin Tone 1 to Tone 2. These were then attached to different types of contexts. The merit of this method is that it allows us to see how different contour tone stimuli might be differentially influenced by various extrinsic contexts. To

address the second question, four types of contexts were used, presented to listeners randomly. We predicted that if general auditory processing played a role in what has been reported for tone normalization processes, then the four types of contexts sharing the same F0 range should elicit qualitatively similar context effects on Mandarin tone perception. If not, domain-specific mechanisms of speech processing would more likely have been involved.

2 Methods

2.1 Participants

Eighteen native Mandarin listeners (nine female, nine male; mean age =25.1 year, SD=1.5) were paid for their participation. Peng et al. [16] reported that different tone inventories affected tone categorical perception without context. Therefore, different tone inventories of participants' dialects may also affect the perception results in the current study. For this reason, to increase participant homogeneity we recruited only right-handed participants from Northern China who only knew Mandarin (and no other Chinese dialects). None of the participants had formal musical training, or any speech, language, or hearing difficulty. All participants gave informed consent in compliance with the protocols approved by the Behavioral Research Ethics Committee of SIAT (Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences).

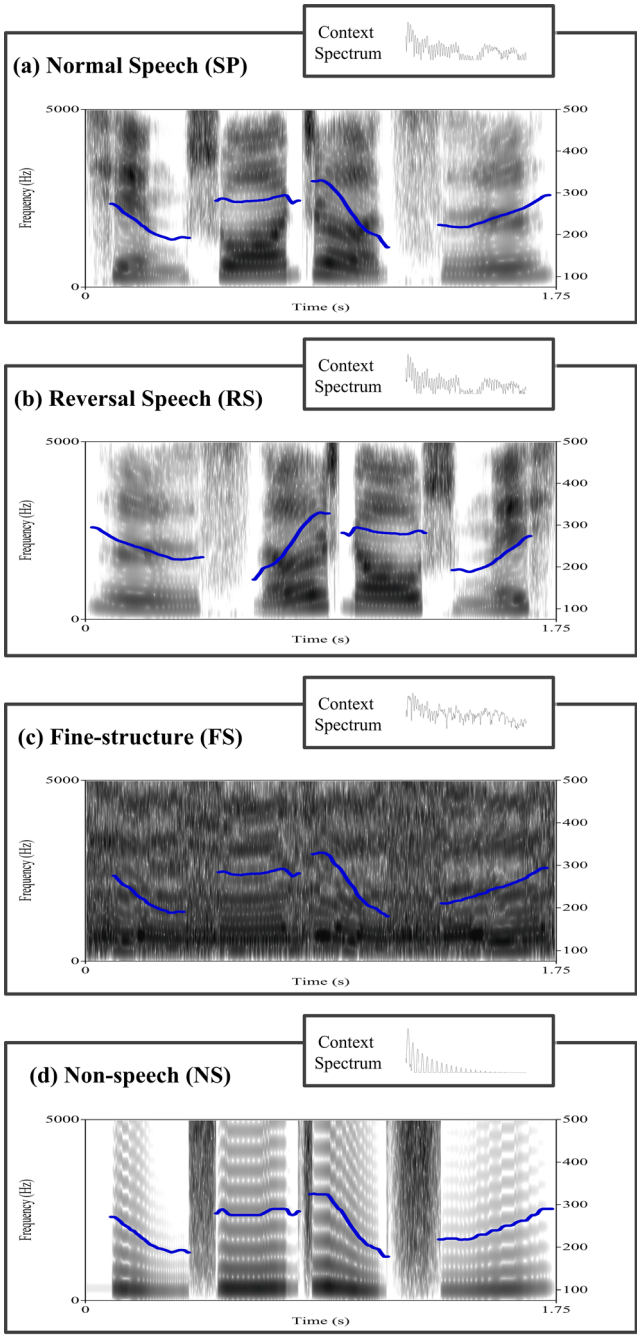
2.2 Stimuli

The context stimuli were generated based on a clear recording of a female native Mandarin talker (using E-MU 0404 USB 2.0, 22050 Hz sampling rate, 16 bit resolution) uttering the Mandarin sentence: 请说这词 /qing3 shuo1 zhe4 ci2/ (please say this word) [6]. This semantically-neutral sentence was chosen because it contains all four Mandarin tones.

Among three recorded tokens of the context utterances, a single utterance was chosen based on its clarity. This original utterance had a mean F0 of 218 Hz with a range of 137–295 Hz. Two versions of the utterance were then created by shifting the entire F0 contour such that the average F0 of the high-frequency context was 250 Hz, and the average F0 of low-frequency one was 190 Hz (using PRAAT 4.0) [17], both with a duration of 1750 ms. These two mean F0 frequencies (250 and 190 Hz) corresponded to the onset F0 values measured from the recordings of the same talker uttering Tone 1 and Tone 2 syllables. Studies, e.g., [18] found that temporal envelope is critical for speech perception whereas temporal fine structure is critical for pitch perception. Fine-structure sound (FS) was made on the basis of the normal speech. Any sound can be mathematically factored into the product of a slowly varying envelope (also called modulation), and a rapidly-varying FS (also known as carrier) [19]. The pitch information is exclusively represented by the fine structure. As for more detailed description of the FS, and the method to construct it, please refer to the following website: https://research.meei.harvard.edu/chimera/index.html. The software we used to construct our FS materials was downloaded from the above website. Reversal speech (RS) was produced by inverting the normal speech on a time scale. Afterwards, the F0 trajectories of normal speech contexts were extracted to synthesize non-speech (NS) contexts with equal-amplitude triangle waves, which (as pointed out above) have a different harmonic structure from that of speech sounds. Each triangle wave was modeled after its speech counterpart: as for a voiced portion, a series of triangular waves were used to model those F0 periods successively; the intensity level of each triangular period was equal to its corresponding speech period. As for an unvoiced portion, it was first divided into smaller segments according to the average F0 values near its both ends (the last F0 value of its immediately preceding voiced portion, and the first F0 value of its immediately following voiced portion); a series of segments of white noise was used to model the unvoiced portion; the intensity level of each noise segment was equal to its corresponding speech segment. Finally, the overall intensity level was raised by 20 dB. The reversal speech, fine-structure sound and the non-speech are all unintelligible to listeners. The four types of contexts share the same F0 profile, for both high-frequency and low-frequency contexts. Altogether, manipulation of contextual F0 height gave rise to eight context conditions (four types of context (SP, RS, FS, NS)×2 context frequency (high, low)). Figure 2(a), (b), (c) and (d) show the representative spectrograms of four types of contexts with the high F0 condition. The non-speech context sounds lower perceptually when compared to the speech context of the same intensity. For the purpose of matching the loudness level of the four contexts, the average intensity level of non-speech contexts was set to 80 dB, 20 dB higher than the other three contexts. It was rated by the authors and two native listeners that the non-speech contexts (80 dB) sounded similar in loudness level as the other three contexts (60 dB). Moreover, the overall shape of the intensity profile of the non-speech stimuli was closely matched to that of other three contexts.

The Mandarin syllable /i/ was chosen as the target syllable, which was derived from natural recordings of the same female talker from whom the context utterances were obtained. In Mandarin, the syllable /i/ when spoken with high-level tone means "clothes", and was coded as stimulus Number 11. However, when /i/ is spoken with the mid-rising tone, it means "aunt", and this was coded as stimulus Number 1.

The major procedures for synthesizing the stimuli were: (1) Adjusting the duration of the target syllable to 450 ms, and fixing the pitch contour to a high level frequency of 250Hz;



◀ **Figure 2** The representative spectrograms in time × frequency scales for higher mean F0 conditions of SP, RS, FS and NS were shown in (**a**), (**b**), (**c**) and (**d**) respectively, with *blue lines* representing F0 contour. The *insets* in panels (**a**)–(**d**) illustrate the spectral slice of the context sounds. The target tone continuum was shown in (**e**).

(2) reducing the number of pitch points to 3, with one at the starting position, one at the 90 ms position, and one at the end position; (3) synthesizing the tone continuum by dragging the aforementioned three pitch points accordingly. The different starting frequencies for the target stimuli were determined by the formula "190Hz+6Hz×(Stimulus Number - 1)".

Context sounds and the target syllables were paired in a random order, with an interval of 200 ms. Figure 3 shows the schematic illustration of stimuli presentation.

2.3 Procedure

A practice block including the four types of contexts was presented to familiarize the subjects with the experimental procedure. Context sounds and target tones were randomly paired, giving rise to 704 trials in all (four types of contexts×2 context frequency×11 target stimuli, repeated eight times), which were then further divided into four blocks. The 11 stimuli were repeated twice in each block. The participants were asked to perform a two-alternative forced choice (2AFC) identification task. They were explicitly asked to pay attention to the whole utterance and then identify the target word by pressing keyboard buttons of "1" (Tone 1 syllable) or "2" (Tone 2 syllable) respectively using the right hand. The whole experiment, including the practice block, was controlled using E-prime, taking approximately 1 h to finish.

3 Results

For a particular stimulus, the identification score was defined as the percentage of responses with which participants identified that stimulus as being either 'Tone 1' or 'Tone 2'. Figure 4 displays only the mean percentage of Tone 2 responses in two F0 shift conditions (high vs. low). Figure 4(a) illustrates the results under normal speech (SP) context, Fig. 4(b) under reversal speech (RS) context, Fig. 4(c) under fine-structure sound (FS) context, and Figure (d) under non-speech (NS) context. The solid line illustrates responses to the target stimuli in the high-F0 context, whereas the dashed line represents those in the low-F0 context.

To give a clearer overview of the data, and to estimate the identification boundaries and shifts as a function of the preceding context, results shown in Fig. 4 were reanalyzed using the Probit analysis, which is designed specifically for the estimation of discrete decision variables [20]. To calculate the boundaries, a cumulative normal curve was also used by transforming the percentage of "Tone 2" responses into z-scores and by finding the best fitting line via linear regression.

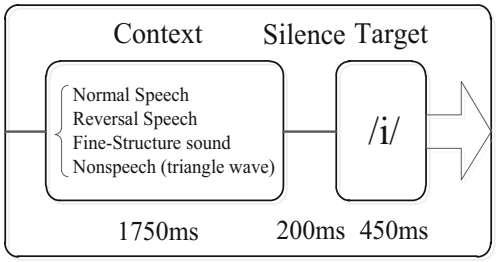


Figure 3 Schematic illustrations of stimuli presentation.

Consistent with previous literature, the boundary of categorical tone perception was taken to be the onset F0 of the target stimuli corresponding to 50 % on this line [16]. Detailed results are given in Table 1 and Fig. 5.

A 4 (context type: SP, RS, FS, NS)×2 (context frequency: high, low) repeated measures ANOVA was conducted on the categorical boundary values, with context type and context frequency as two within–subjects factors. Where appropriate, the Greenhouse-Geisser method was used to correct violations of sphericity. The analysis confirmed a significant main effect of context type, F (3, 51)=3.215, p<0.05, with no significant main effect being found for the context frequency, F (1, 17)= 1.205, p=0.288. There was a significant interaction between context type and context frequency, F (3, 51)=2.948, p<0.05.

After this, simple main effect analyses of the context frequency were conducted with Bonferroni adjustment. For the normal speech condition (SP), context frequency (high vs. low) showed a significant effect (F (1, 17)=10.47, p<0.01), leading to an approximately 0.27 difference in categorical boundary positions (see Table 1). But no such effects were found for the context of reversal speech (F (1, 17)=0.106, p=0.749), fine-structure sound (F (1, 17)=0.821,p=0.377), or non-speech (F (1, 17)=0.074,p=0.788). Thus, it appeared that the main effect of context type was driven largely by the significant effect of the normal speech context, and likely by the perception of ambiguous tone stimuli in the middle portion of the continuum (see Fig. 4, from No. 5 to No. 8 stimuli).

In addition, the effects of high-F0 and low-F0 conditions on No. 7 stimulus were somewhat different in the non-speech context [see Fig. 4(d)]. But paired samples T-test revealed no significant difference between F0 conditions in the non-speech context for the most affected No. 7 stimulus (t (17)= 1.948, p=0.068). Likewise, the effects of context frequency fell short of significance in the reversal and fine-structure sound contexts.

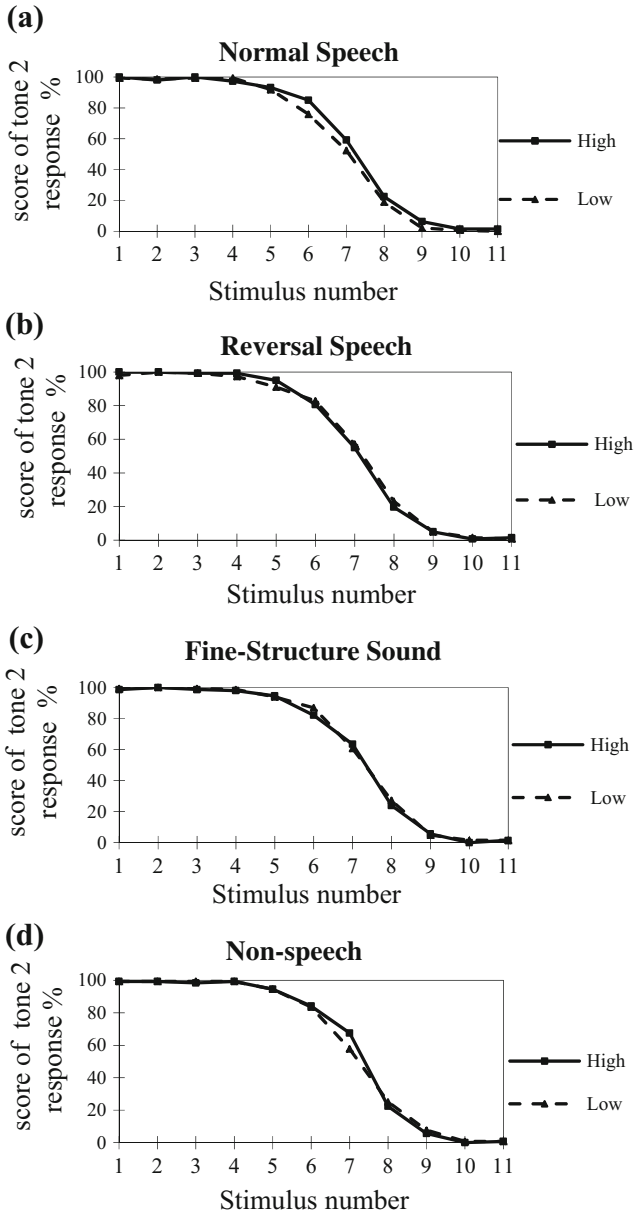


Figure 4 Mean percentage of Tone 2 responses for the four context conditions. Responses to the high and low F0 conditions were shown respectively by *solid* and *dashed lines*.

4 Discussion

4.1 The Effects of Extrinsic Contexts on the Perception of Mandarin Tone

Previous research has provided strong evidence for the context-dependency of lexical tone perception, notably for the perception of level tones that possess very similar F0

Table 1 Derived categorical boundary positions for each type of context with high and low mean F0.

Contexts	High	Low	Difference
Normal speech	7.13	6.86	0.27
Reversal speech	7.07	7.03	0.04
Fine-structure sound	7.14	7.22	0.08
Non-speech	7.19	7.16	0.03

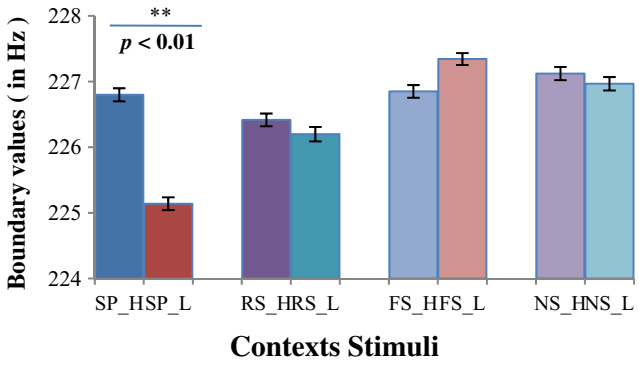


Figure 5 Categorical boundary values (represented by starting F0 in Hz which is determined by the formula "190Hz+6Hz×(categorical boundary position - 1)") with four types of contexts ("H" means the higher mean F0 in each context, while "L" represents the lower mean F0 in each context). A higher F0 boundary value indicates more Tone 2 responses.

contours [3–5], but results have been mixed for lexical tones differing along both F0 height and direction dimensions [6–10]. Explanations for such mixed results are twofold. On the one hand, Mandarin tone perception may rely on word-internal F0 characteristics more than external ones, making it less susceptible to the influence of the preceding speech context [10]. Such a finding has been recently corroborated by Peng et al. [21], who found that native Mandarin subjects, in contrast to Cantonese subjects, were able to show stable talker normalization even in the absence of contextual cues. On the other hand, there is ample literature showing that Mandarin listeners assign different weights to the perceptual dimensions of Mandarin tones. In a seminal study on tone perception, Gandour [22] observed that pitch direction tends to outweigh height in terms of its perceptual importance. In another study, Li and Zhang [23] reported that it is possible for Mandarin listeners to identify Tone 1 as opposed to Tone 4 based solely on F0 slope, and not on F0 onsets. Though differing in approaches, these studies converge on the idea that extrinsic context is relatively less important than intrinsic cues for resolving lexical tone ambiguity in Mandarin, and that height is less crucial than direction.

The above viewpoints are also corroborated by our study. The effect of normal speech context observed for Mandarin tones in this research was much more restricted than those observed for Cantonese level tones. Whereas the perception of Cantonese level tones can be highly predictable based on the mean F0 of the surrounding speech context [3–5], the same cannot be said about Mandarin tone perception. In our study, even for the most affected No. 6 stimulus (onset F0=220 Hz), the percentage of Tone 2 responses improved by only 10 %: from 75 % with lower mean F0 context to 85 % with higher mean F0 context. There was also barely any effect on the unambiguous tone stimuli close to the two ends of the tone continuum.

Even though the influence of speech contexts on Mandarin tone perception is highly restricted, the overall trend of the context effect is still in a contrastive manner. Consequently, when preceded by a high-frequency context, the perception of onset F0 was lowered for the target tone, leading to a greater proportion of mid-rising tone responses rather than high-level tone responses. This is consistent with the apparent influence of context on phonetic categorization [11–14].

4.2 The Potential Mechanisms of Lexical Tone Normalization

Previous studies diverge on the kind of processing mechanisms involved in lexical tone normalization. Some suggest that such processes recruit speech-specific mechanisms [3, 5], whereas others suggest that the normalization of lexical tones is mediated by the general perceptual mechanisms [6, 15]. Given the current experimental results, our research supports the speech-specific mechanisms underlying lexical tone normalization.

Consistent with Zhang et al. [5] where speech and non-speech contexts were found to exert unequal effects on the perception of Cantonese level tones, the context effect that surfaced in our study was also driven mostly by the speech context, with the identification preference being modified only mildly by the non-speech contexts (see Fig. 5). These results support the idea that tone normalization is the result of speech-specific mechanisms. Moreover, the context of fine-structure sound, which makes use of a more complicated spectral composition, also exerted no obvious context effects on target tone perception in the present study. As for the reversal speech context, it is more comparable with the meaningless [ ] context used in Francis et al. [3]. In both studies, the unintelligible contexts fall short of facilitating normalization presumably because the native Cantonese listeners seemed to be either unable or unwilling to use meaningless contexts for the purposes of lexical tone normalization. It thus seems plausible that some properties of the unintelligible context failed to match sufficiently with the corresponding property of the target syllable, leading listeners to dissociate (consciously or unconsciously) the two types of signals and engage distinct processing mechanisms by only considering internal F0 cues in lexical tone judgments.

Zhang et al. [24] also reported that nonsense speech (although each syllable remained a meaningful morpheme) showed a similar context effect as did meaningful speech. This suggests the possibility that talker normalization can be facilitated by the presence of familiar phonological cues alone in the context, regardless of whether the semantic content is meaningful or not. For native Mandarin listeners, the reversal speech used in the current study sounds like unintelligible foreign speech, which is phonetically possible but phonologically unfamiliar. Therefore, to activate lexical tone normalization that may operate with speech-specific mechanisms, the speech contexts need to be phonologically meaningful, for

both level and contour tones. In other words, native listeners need to be able to classify phonetic units within the utterances into familiar phonological categories. Furthermore, the so-called "cocktail party" phenomenon [25], where reliable talker normalization requires the ability to accurately tune to a target talker's voice while filtering out irrelevant voices and unintelligible non-speech sounds, may share the same mechanisms as lexical tone normalization.

The existence of speech-specific mechanisms is further suggested by an event-related potential (ERP) study of Zhang et al. [24], who examined the time course of context-dependent talker normalization in spoken word identification. Compared with a non-speech context, the speech contexts enabled listeners to tune to a talker's pitch range. In this way, it is the speech contexts that induce more efficient talker normalization during the activation of potential lexical candidates and lead to more accurate selection of the intended word in spoken word identification.

However, how can the contradictory conclusions by Huang and Holt [6], advocating the general perceptual mechanisms in tone normalization, be explained? In their study, non-speech contexts modeling the mean spectrum of F0 also elicited a significant effect on the perception of contour tones. One reason could be attributed to the differences in the non-speech materials. In the study of Huang and Holt [6], the materials of the non-speech context used were a sequence of pure tones composed of four equal-amplitude sine-waves with frequencies at the first four multiples of the F0 in Experiment 2, and only one single sine-wave at the frequency of the first harmonic in Experiment 3. However, the material of non-speech context used in our research is the triangle wave, which is a non-sinusoidal waveform named for its triangular shape. Like a square wave, a triangle wave contains all the odd harmonics. While both types of non-speech materials retain the complete information of F0, there are much richer harmonics in triangle waves than in the sine-wave tones. Therefore, our results suggest that Huang and Holt's finding regarding contextual effect of their non-speech contexts is not generalizable to all types of non-speech contexts. The other reason may have been because the same group of subjects participated in the Experiment 1 of speech context first, and were then later recruited for the second and third experiments of non-speech contexts just 10 days later. Since the speech and non-speech contexts shared exactly the same F0 contour, there may have been a carryover of a facilitation effect from their speech vs. non-speech contexts.

5 Conclusions and Future Work

Native Mandarin listeners' categorical perception of Mandarin Tone 1 and Tone 2 was investigated with different preceding contexts. Results show that the Mandarin tones are influenced in a contrastive manner only with normal speech context, affecting mainly the middle part of the tone continuum where the sounds are also the most perceptually ambiguous. Moreover, the results suggest that Mandarin tone normalization recruits speech-specific mechanisms and the phonological categories in the speech context need to be understood by native Mandarin listeners.

There is still much to be understood about tone perception and normalization, especially with respect to possible similarities and dissimilarities between the processing mechanisms used for speech and non-speech more generally. Besides Mandarin Tone 1 and Tone 2, the context effect on some other tone categories in Mandarin could also be studied in the future. In addition, the high-level tone and the mid-rising tone examined in this study exist in both the Mandarin and Cantonese tone inventories, so the question of whether native Cantonese listeners show the same contextual dependence of tone perception merits future study.

Acknowledgments This work was partially supported by grants from National Natural Science Foundation of China (NSFC: 61135003 and NSFC: 11474300). We are thankful to all the subjects at SIAT who participated in this experiment. We thank Professor Feng Shi at Nan Kai University for his constructive comments as this work progressed.

References

1. Wang, W. S.-Y. (1972). The many uses of F0. In A. Valdman (Ed.), *Linguistics and phonetics to the memory of Pierre Delattre* (pp. 487–503). The Hague: Mouton.
2. Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: a corpus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics, 34*(1), 134–154.
3. Francis, A., Ciocca, V., Wong, N., Leung, W., & Chu, P. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America, 119*, 1712–1726.
4. Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intra-talker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46*, 413–421.
5. Zhang, C. C., Peng, G., & Wang, W. S.-Y. (2012). Unequal effects of speech and non-speech contexts on the perceptual normalization of Cantonese level tones. *Journal of the Acoustical Society of America, 132*, 1088–1099.
6. Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America, 125*, 3983–3994.
7. Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics, 11*, 373–382.
8. Fox, R., & Qi, Y.-Y. (1990). Context effects in the perception of lexical tone. *Journal of Chinese Linguistics, 18*, 261–283.
9. Lin, T., & Wang, W. S.-Y. (1984). Shengdiao ganzhi wenti. *Journal of Chinese Linguistics, 2*, 59–69.
10. Moore, C., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America, 102*, 1864–1877.
11. Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.
12. Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*, 407–412.

13. Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America, 102*, 1134–1140.
14. Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America, 108*, 710–722.
15. Huang, J., & Holt, L. L. (2011). Evidence for central origin of lexical tone normalization. *Journal of the Acoustical Society of America, 129*, 1145–1148.
16. Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics, 38*, 616–624.
17. Boersma, P. & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 4.0). http://www.praat.org/, last accessed 14 Nov 2014.
18. Zeng, F.-G., Nie, K.-B., Liu, S., Stickney, G., Del Rio, E., Kong, Y.-Y., & Chen, H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues. *Journal of the Acoustical Society of America, 116*, 1351–1354.
19. Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*, 87–90.
20. Finney, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.
21. Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. M., & Wang, W. S.-Y. (2012). The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research, 55*, 579–595.
22. Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11*, 149–175.
23. Li, B. & Zhang, C.C. (2010). Effects of F0 dimensions in perception of Mandarin tones. *The 7th International Symposium on Chinese Spoken Language Processing, 29 November – 02 December 2010, National Cheng Kung University*, Tainan, Taiwan.
24. Zhang, C. C., Peng, G., & Wang, W. S.-Y. (2013). Achieving constancy in spoken word identification: time course of talker normalization. *Brain and Language, 126*, 193–202.
25. Bronkhorst, A. W. (2000). The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica, 86*, 117–128.

Fei Chen He received his M.S. degree in 2014 in Linguistics from Nankai University, Tianjin, China. He is now working as a Research Assistant at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His current research interests include lexical tone normalization, the development of tone perception, and pathological linguistics.

Gang Peng He received his Ph.D. in Language Engineering from City University of Hong Kong in 2002. He has published several research articles in various high-profile international journals. He is now a Research Associate Professor at Department of Linguistics and Modern Languages in Chinese University of Hong Kong (CUHK), and Adjunct Professor of Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences. He is also Deputy Director of the Chinese University of Hong Kong (CUHK)-Peking University (PKU)-University System of Taiwan (UST) Joint Research Centre for Language and Human Complexity. His central research focus is on how language is represented and processed in the human brain, and how different cultures, reflected in their languages, shape perception differently. His research areas include psycholinguistics, neurolinguistics, experimental phonetics, computational/corpus linguistics, hearing disorders, and related topics.