Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism

Kaile Zhang, Xiao Wang and Gang Peng

Citation: The Journal of the Acoustical Society of America **141**, 38 (2017); doi: 10.1121/1.4973414
View online: https://doi.org/10.1121/1.4973414
View Table of Contents: https://asa.scitation.org/toc/jas/141/1
Published by the Acoustical Society of America

Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism

Kaile Zhang,[1] Xiao Wang,[2] and Gang Peng[1,a)]
[1]*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China*
[2]*Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China*

(Received 23 April 2016; revised 19 November 2016; accepted 8 December 2016; published online 4 January 2017)

Context is indispensable for accurate tone perception, especially when the target tone system is as complex as that of Cantonese. However, not all contexts are equally beneficial. Speech contexts are usually more effective in improving lexical tone identification than nonspeech contexts matched in pitch information. Some potential factors which may contribute to these unequal effects have been proposed but, thus far, their plausibility remains unclear. To shed light on this issue, the present study compares the perception of lexical tones and their nonlinguistic counterparts under specific contextual (speech, nonspeech) and attentional (with/without focal attention) conditions. The results reveal a prominent congruency effect—target sounds tend to be identified more accurately when embedded in contexts of the same nature (speech/nonspeech). This finding suggests that speech and nonspeech sounds are partly processed by domain-specific mechanisms and that information from the same domain can be integrated more effectively than that from different domains. Therefore, domain-specific processing of speech could be the most likely cause of the unequal context effect. Moreover, focal attention is not a prerequisite for extracting contextual cues from speech and nonspeech during perceptual normalization. This finding implies that context encoding is highly automatic for native listeners. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4973414]

[JFL] Pages: 38–49

I. INTRODUCTION

A. Talker variability and lexical tone normalization

The acoustic realization of speech varies dramatically across talkers, as talkers differ in both speaking styles and the configurations of their vocal tracts (Peterson and Barney, 1952). Even the acoustic characteristics of one talker's speech change rather drastically under different physical and psychological conditions (Garrett and Healey, 1987; Protopapas and Lieberman, 1997). It is therefore natural to ask how speech categories are represented in the human brain in the face of speech variations. Several studies (e.g., Gerstman, 1968; Syrdal and Gopal, 1986) have argued that each phonological category has only one abstract mental representation. To correctly categorize incoming speech signals, listeners need to be able to compensate for individual differences in speech production and to reconstruct the intended phonemic target. Such a process, by which multiple acoustic variants are mapped onto the same phonological category, has been known as perceptual normalization (Johnson and Mullennix, 1997). This normalization process applies not only to phonetic segments, such as vowels and consonants, but also to suprasegmental components, such as lexical tones. Since lexical tones are used to distinguish lexical meanings (Wang, 1972), the normalization of lexical tones is of great importance for tone language speakers.

The normalization of lexical tones is frequently found to rely on both intrinsic and extrinsic cues. Intrinsic cues refer specifically to the acoustic correlates of the target words, including fundamental frequency (*F*0), intensity, duration, and voice quality (Zhang *et al.*, 2012). While all of these are potentially useful, *F*0 seems to be the primary cue to categorize lexical tones (Wang, 1972; Bishop and Keating, 2012; Zhang *et al.*, 2012). External cues, on the other hand, refer to the information provided by surrounding contexts. Both the *F*0 ranges and the average *F*0 heights of the contexts are useful for lexical tone perception (Leather, 1983; Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al.*, 2006). However, *F*0 ranges appear to be the more effective cues, compared to mean *F*0 heights, in highly ambiguous situations (Zhang *et al.*, 2012).

B. Context effects in lexical tone normalization

Lexical tones can be divided into two broad types: contour tones and level tones (Pike, 1948). Contour tones are characterized by distinctive pitch shapes, changing their pitch heights over the time course of syllables, whereas the heights of level tones remain relatively steady (Yip, 2002). According to this criterion, Cantonese has three contour tones (high rising /25/, low rising /23/, and low falling /21/)

[a)]Also at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Boulevard, Shenzhen, 518055, China. Electronic mail: gpengjack@gmail.com

and three level tones (high-level /55/, mid-level /33/, and low-level /22/). Many studies (e.g., Huang and Holt, 2009, 2011; Peng *et al.*, 2012) have shown that, for contour tones, *F*0 dynamics alone are sufficient to differentiate tone categories. For level tones, however, the situation is remarkably different.

On the one hand, *F*0 dynamics can rarely be used to distinguish between level tones, due to the similarity of their pitch contours (Moore and Jongman, 1997; Francis *et al.*, 2006). On the other hand, ambiguity caused by inter- and intra-talker variability makes absolute pitch heights a less effective cue for lexical tone perception (Wong and Diehl, 2003; Francis *et al.*, 2006; Peng *et al.*, 2012). Therefore, distinguishing Cantonese level tones by intrinsic cues alone poses a serious challenge. Similar observations have been reported in Peng *et al.* (2012), in which native Cantonese speakers were asked to classify isolated pitch stimuli as one of the six tones in Cantonese. In that study, targets to be identified were synthesized based on speech samples produced by native Cantonese speakers with different average pitch heights. It was found that, when presented in isolation, the perceptual mapping of Cantonese level tones overlapped extensively across tone categories. However, when similar pitch targets were embedded in speech contexts, their identification rates improved significantly in a study using essentially the same methods to introduce talker variability (Zhang *et al.*, 2012).

A noteworthy characteristics of the normalization process is that contexts typically exert a contrastive effect on the perception of Cantonese level tones (Wong and Diehl, 2003; Francis *et al.*, 2006; Zhang *et al.*, 2012). That is, a word carrying a mid-level tone is perceived as having a high-level tone in contexts with low average *F*0's and as having a low-level tone in contexts with high average *F*0's. This might be caused by the way in which listeners normalize tones with extrinsic context cues. According to the *F*0 range assessment model (Wong and Diehl, 2003; Francis *et al.*, 2006), extrinsic tone normalization is complex and involves a series of steps. Listeners might extract the useful context cues first. After perceiving the target sound, they integrate these two parts of information together and calculate the relative pitch heights of the targets against talker-specific *F*0 ranges estimated from the immediate context. Broadly speaking, the assumptions of the *F*0 range assessment model are consistent with the context tuning mechanism proposed by Joos (1948), as both of them highlight the role of contextual information and target-context integration in perceptual normalization.

C. Speech-specific context effects

Previous studies (Francis *et al.*, 2006; Zhang *et al.*, 2012, 2013) have further compared different types of contexts and found that only speech contexts significantly improve the normalization of lexical tones. In Francis *et al.* (2006), the authors created an unintelligible context by extracting the original *F*0 contour of a Cantonese sentence and re-synthesizing it with the "hummed" neutral vocal tract function in Praat (Boersma and Weenick, 2012). The results show that listeners recognized the target tones primarily based on their absolute pitch heights. While such a pattern could indeed suggest that linguistically meaningless contexts play a limited role in tone normalization, results obtained in Francis *et al.* (2006) were inconclusive, for the context was highly perceptually ambiguous. In fact, the synthesized context was acoustically similar to the schwa [ə] in terms of formant distribution, which meant that listeners could perceive it as either speech or nonspeech. Since Francis *et al.* (2006) reported that Cantonese listeners could normalize Cantonese tones with the help of English contexts, they deduced that the schwa-like context was most likely to be perceived as nonspeech. An alternative explanation for this is that the schwa-like context was ignored by listeners because it was too "foreign" to be relevant to the lexical tone judgment (Francis *et al.*, 2006). To avoid this dilemma, Zhang *et al.* (2012, 2013) synthesized the nonspeech analogs of the speech contexts using triangle waves. Their results showed that, whereas listeners encountered obvious difficulties in identifying lexical tones embedded in nonspeech contexts, their accuracy improved dramatically when the carrier sentences were normal speech. In Secs. I C 1–I C 3, some potential factors which may cause the superiority of speech contexts in lexical tone normalization are briefly reviewed, followed by the questions to be addressed in the current study.

1. The focal attention

Some studies (e.g., Francis *et al.*, 2006; Zhang *et al.*, 2012) proposed that the unequal context effect is due to the absence of the focal attention on the nonspeech contexts. Listeners in previous studies (e.g., Francis *et al.*, 2006; Zhang *et al.*, 2012) were usually required to indicate their perceptual judgments by performing a word identification task (i.e., choosing the word that has the same pronunciation as the probe they had just heard). The linguistic nature of the word identification task might have misled participants into believing that only the speech contexts were useful. Nonspeech contexts, on the contrary, were regarded as irrelevant to the linguistic tasks at hand and thus were allocated with little focal attention.

This assumption indicates that the focal attention is indispensable in the context encoding stage of lexical tone normalization and without focal attention neither speech nor nonspeech contextual information can be effectively processed. The focal attention as used here refers to the type of attention being deployed when an individual deliberately and consciously focuses on a task. When the attentional demands of the secondary task increased, participants' reaction times were significantly prolonged in normalizing the consonants of CV syllables (Nusbaum and Morin, 1992). Hugdahl *et al.* (2003) found that focal attention triggered a stronger neuronal activation in perceiving words and vowels, compared to the attention-absent condition. These results are consistent with the view that normalization with extrinsic context cues is constrained by the amount of attentional resources and that focal attention may facilitate speech

perception in general (Nusbaum and Morin, 1992; Hugdahl *et al.*, 2003).

2. The degree of familiarity

Lee *et al.* (1996) believed that Cantonese speakers were more familiar with the lexical tones, whereas nonlexical pitches were new to them and this might be the reason why Cantonese speakers could successfully distinguish lexical tones but not the closely-matched pitches embedded in pseudo words. Such an assumption could also be applicable to tone normalization with extrinsic cues. Since the participants have seldom heard artificially-synthesized sounds, they lacked sufficient familiarity with nonlinguistic stimuli. Naturally, normalization performance was inferior when nonlinguistic contexts were presented instead of speech precursors. Language exposure is an effective way to improve subjects' perception of the unfamiliar languages (Wang *et al.*, 1999, 2003; Wayland and Guion, 2004). Some neurobiological changes also emerged, together with the improvement of speech perception. As reported by Kaan *et al.* (2007), after a two-day perceptual training on Thai tones, the amplitude of the mismatch negativity elicited by English speakers became larger and the later negativity (350–650 ms) elicited by Chinese speakers decreased. Therefore, it is reasonable to assume that daily exposure to speech makes the effects of speech contexts more prominent than nonspeech contexts.

3. The speech-specific mechanism

As suggested by the name, the speech-specific mechanism suggests that there is a network in our brain dedicated to speech signals, which gives rise to domain-specific cognitive processes of speech (Liberman *et al.*, 1967; Liberman and Mattingly, 1985). Zhang *et al.* (2012) also believed that speech is processed by the domain-specific mechanism, which was why the speech targets in their research were barely affected by nonspeech contexts. Many studies have shown the cognitive differences between speech and nonspeech perception. For example, Diehl and Kluender (1989) and Bregman (1990) reported that, after being analyzed in the primary auditory cortex, speech and nonspeech signals are submitted to different cortical areas. Speech signals are sent to the auditory association cortex to match the speech templates formed by long-term linguistic experience and to be processed on a more elaborate level. Apparently, such processes do not apply to nonspeech signals that lack mental representations. Distinct neural processing of speech and nonspeech signals was even observed as early as when the signals reached the primary auditory cortex (the earliest cortical level of sound processing; Whalen *et al.*, 2006). Fedorenko *et al.* (2011) localized the brain regions involved in linguistic and nonlinguistic tasks, and found that regions activated by linguistic tasks (e.g., left inferior frontal gyrus, left temporal structures) showed little response to nonlinguistic functions. Such neurophysiological evidence highlights the functional specificity of the language-processing regions in the brain, thereby offering additional support to the existence of domain-specific mechanisms and the plausibility of the speech-specific normalization.

In summary, Francis *et al.* (2006) and Zhang *et al.* (2012, 2013) reported that speech contexts facilitated listeners' lexical tone normalization more effectively than nonspeech contexts containing the same pitch information. Some potential factors which may cause the unequal context effect have also been highlighted by previous studies. However, whether or not these explanations are empirically valid and how these potential factors interact with each other are still unclear.

D. Research aims

The present study is designed to extend previous research on lexical tone normalization by exploring the causes of the unequal contributions of speech and nonspeech contexts. Specifically, three potential factors, the focal attention, the degree of familiarity, and the speech-specific mechanism, will be tested to see how they contribute to the unequal effect of speech and nonspeech contexts.

II. METHOD

Following previous research (e.g., Wong and Diehl, 2003; Francis *et al.*, 2006; Zhang *et al.*, 2012, 2013), the contrastive context effect was used to index the magnitudes of pitch normalization. If a mid-level pitch could be recognized as a high-level pitch in contexts with low average pitch heights, and as a low-level pitch in contexts with high average pitch heights, then context-dependent perceptual normalization has successfully taken place.

Two experiments were conducted. In the word identification task (experiment I), subjects were asked to normalize lexical tones in both single- and dual-task paradigms (see Fig. 1), in order to explore the relationship between focal attention and the speech-specific context effect. Subjects in the dual-task paradigm were instructed to pay focal attention to the visual task (the secondary task) while passively perceiving the auditory contexts of the primary task. Therefore,

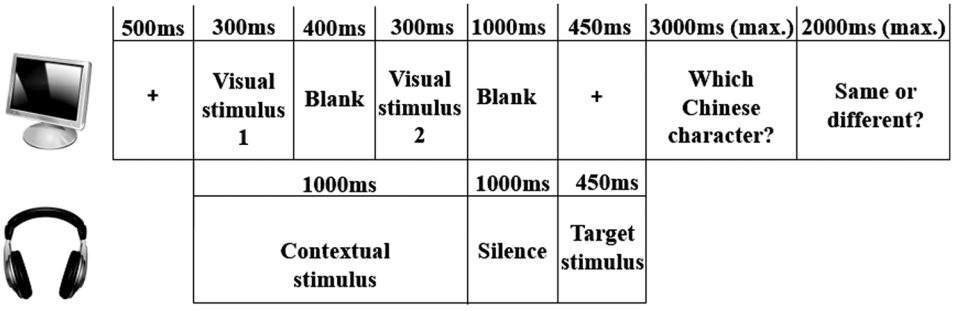

FIG. 1. A schematic illustration of the procedures of the dual-task paradigm.

the attentional resources used in perceiving either speech or nonspeech contextual information, to a large extent, was distracted by the processing of the visual stimuli. If this manipulation reduces the effectiveness of the speech contexts compared with the single-task paradigm, the focal attention is likely to play an important role in the context processing of tone normalization.

If the normalization accuracy is positively related to listeners' familiarity level with the testing stimuli, the identification of nonlinguistic pitches should also be superior when it is preceded by speech contexts with which native speakers are more familiar. Therefore, experiment II employed a pitch location judgment task to test to what extent the unequal context effects are affected by familiarity. In this task, listeners used "high," "middle," or "low" to indicate the heights of nonlinguistic pitch targets, with reference to the preceding speech or nonspeech contexts. Since the actual effect of contextual familiarity may be modulated by the focal attention, experiment II likewise adopted the dual-task paradigm. Additionally, to match the nonlinguistic nature of the pitch location judgment task, experiment II used picture discriminations (rather than homophone judgments) in the secondary task.

The combined results of experiments I and II have the potential to shed further light on the plausibility of the speech-specific mechanism. If the perception of lexical tones and that of nonlinguistic pitches show the same pattern, such results will lend support to the employment of the general perceptual mechanisms in perceptual normalization (Huang and Holt, 2009, 2011).

A. Experiment I: Word identification task

1. Participants

A total of 18 native speakers of Hong Kong Cantonese (nine males), aged 19–23 yrs [mean age = 20.89 yrs, standard deviation (SD) = 1.4], were paid to participate in this experiment. All of them were right-handed, with normal hearing and normal or correct-to-normal vision. No participants had received any long-term professional training in linguistics, psychology, or music. The experiment was approved by the Joint Chinese University of Hong Kong—New Territories East Cluster Clinical Research Ethics Committee. All participants gave informed written consent before the experiment.

2. Stimuli

a. Auditory stimuli. The auditory materials used in Zhang *et al.* (2013) were adopted for the present study. In that study, four native Cantonese speakers (two males) were invited to record the context and target stimuli. Each speaker had a distinct average pitch height [female high (FH): 246 Hz; female low (FL): 210 Hz; male high (MH): 143 Hz; male low (ML): 123 Hz] and a unique though partly overlapping pitch range (FH: 198–294 Hz; FL: 166–279 Hz; MH: 112–194 Hz; ML: 96–151 Hz). This was to introduce inter-talker variability and to elicit the phenomenon of perceptual normalization. The context sentence ["呢個字係" (/li55 ko33 tsi22 hɐi22/ "This word is…")] and the target word ["意" (/ji33/ "meaning")] were read by each speaker 6 times. Among the six recordings, the recording with a mean *F*0 closest to the grand mean *F*0 of the six recordings was chosen as the stimulus. The duration of the context stimuli was normalized to 1000 ms. The same procedures were used to select and normalize filler sentences "請留心聽" (/tsʰiŋ25 lɐu21 sɐm55 tʰiŋ55/ "Please carefully listen to…") and "我以家讀" (/ŋo23 ji21 ka55 tuk2/ "Now I will read…").

To trigger the contrastive context effect experimentally, the *F*0 trajectories of the context stimuli (fillers excluded) were shifted either three semitones up or three semitones down from their original heights. To match the preceding contexts, target words "意" (/ji33/ meaning) were adjusted to 55 dB in intensity and 450 ms in duration. They also matched the preceding context with regard to the talker. The nonspeech stimuli were synthesized with triangle waves, closely imitating the *F*0 and the intensity profiles of their speech counterparts. All of the adjustments mentioned above were carried out by Praat, generating 32 auditory stimuli [four talkers (FH, FL, MH, and ML) × three *F*0 shifts (raised, unshifted, and lowered) × two context types (speech context and nonspeech context) + four talkers × one filler × two context types] altogether.

b. Visual stimuli. The stimuli of the secondary task (i.e., homophone judgment) were presented visually. Out of the 192 pairs of traditional Chinese characters, 128 pairs were homophones, whereas the remaining 64 pairs were non-homophones. Each pair of the visual stimuli was presented only once during the whole task. All the visual stimuli were presented in white ink against a black background.

3. Procedures

Two blocks of the lexical tone normalization task, one single- and one dual-task block, were presented in a counterbalanced order across participants. The 32 auditory stimuli were repeated 6 times and were presented to the subjects in a randomized order. Subjects received written instructions prior to each block of experiments.

In each trial of the single-task block, a context stimulus lasting for 1 s was played after the fixation sign "+." After 1000 ms silence, the target stimulus was played, followed by the visual prompt "which Chinese character?" on the screen. After hearing the target, participants needed to identify the words as soon as possible by pressing the buttons labeled "醫" (/ji55/ "a doctor"), "意" (/ji33/ meaning) or "二" (/ji22/ "two") on the computer keyboard. The maximum (max.) allowable response time was 3000 ms.

In each trial of the dual-task block, two traditional Chinese characters were displayed successively, after the fixation sign +, each lasting for 300 ms with a 400 ms inter-stimulus interval. The auditory context stimulus was presented simultaneously. To perceive the Chinese characters successfully in such a short duration as 300 ms (for details, please see Sec. IV), subjects must pay their full attention to the visual task. After an interval of 1000 ms, the target stimulus was played with a fixation sign + shown in the center of the screen. The prompt, which Chinese character?, appeared

on the screen after the target stimuli. Participants were instructed to identify the target word as soon as possible after seeing this prompt. Upon receiving the participants' judgments, another visual prompt "same or different?" was presented, to ask participants to respond to the secondary task by pressing the corresponding buttons. Practice sessions were given before each testing block.

B. Experiment II: Pitch location judgment task

1. Participants

The criteria for selecting subjects were similar to those of experiment I. A total of 15 right-handed native speakers of Hong Kong Cantonese (seven males), aged 18–25 (mean age = 21.2 yrs, SD = 1.93) were paid to participate. None of them took part in experiment I. Prior to the experiment, the subjects gave their written consent in compliance with the experimental protocol approved by the Joint Chinese University of Hong Kong—New Territories East Cluster Clinical Research Ethics Committee.

2. Auditory and visual stimuli

The auditory stimuli functioning as speech contexts and nonspeech contexts were the same as those used in experiment I. Nonspeech targets were synthesized using triangle waves, maximally replicating the pitch trajectories and the intensity profiles of the speech targets. To keep the results of the two experiments comparable, the same context and target manipulations were performed, generating a total of 32 auditory stimuli (four talkers × three *F*0 shifts × two context types + four talkers × one filler × two context types). For the secondary task (picture discrimination), 10 different pictures were created by organizing 14 white squares in various fashions against a black background (see Fig. 2 for sample pictures).

3. Procedures

Similar to experiment I, participants were required to accomplish a single-task block and a dual-task block. The procedures of the single-task blocks were largely identical for experiments I and II, except that the prompt was changed to "which pitch height?" in experiment II. The participants were instructed to judge the relative pitch heights of the targets with reference to the pitch ranges of the preceding contexts. Responses were made by pressing the buttons labeled as "高" (high), "中" (middle), and "低" (low) on the keyboard. The procedures of the dual-task block were largely the same as those described above for experiment I, except that the visual word stimuli were replaced with pictures.

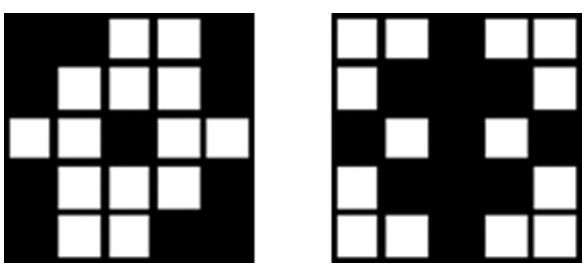

FIG. 2. Sample pictures used in the picture discrimination task of experiment II.

C. Data analysis

The experimental results were analyzed using two different measures, the perceptual height and the identification rate, both of which have been widely used in tone perception studies (Lee *et al.*, 1996; Wong and Diehl, 2003; Francis *et al.*, 2006; Zhang *et al.*, 2012).

A typical production of the Cantonese mid-level tone is about two semitones higher than a low-level tone and three semitones lower than a high-level tone (Chao, 1969). In the current paradigm, "6," "3," and "1" were used to respectively represent the perceptual heights of Cantonese high-, mid-, and low-level tones (Wong and Diehl, 2003; Zhang *et al.*, 2012). Data were coded in similar ways for the pitch location judgment task, with 6 being used as an index for high, 3 for mid, and 1 for low pitch responses. If the perceptual height is close to 6, it means that, due to context manipulations, the target sounds were more frequently perceived as having high-level pitches. Conversely, if the perceptual height is close to 1, it means that low pitch responses were made more frequently than others. Comparatively speaking, a perceptual height close to 3 is much less revealing, as there are too many possibilities behind such a result. To narrow down the potential uncertainty, the present study analyzed participants' identification rates in tandem with their perceptual heights.

Based on the above-noted contrastive context effect, the high-, mid-, and low-level pitch responses were defined as the correct responses in the *F*0-lowered, the *F*0-unshifted, and the *F*0-raised contexts, respectively. The percentage that the target stimuli were identified as the correct responses was calculated as the identification rate. A higher identification rate may indicate a successful perceptual normalization.

III. RESULTS

A. Perceptual height analysis

1. Experiment I: Word identification task

Figure 3 displays the average perceptual height as a function of *F*0 conditions and task paradigms in the word identification task. As can be seen, *F*0 shifts in nonspeech contexts made little difference to perceptual heights, regardless of the task paradigm. By contrast, average perceptual heights were pronouncedly different across the *F*0 conditions in the speech contexts for single- and dual-task blocks alike. The mean perceptual height was close to 1 in the *F*0-raised contexts, 3 in the *F*0-unshifted contexts, and 6 in the *F*0-lowered contexts, consistent with the predictions of the contrastive context effect. These results corroborated the finding that contexts of different natures (speech/nonspeech) contribute unequally to lexical tone normalization, with the effectiveness of speech contexts significantly surpassing that of nonspeech contexts (e.g., Zhang *et al.*, 2012).

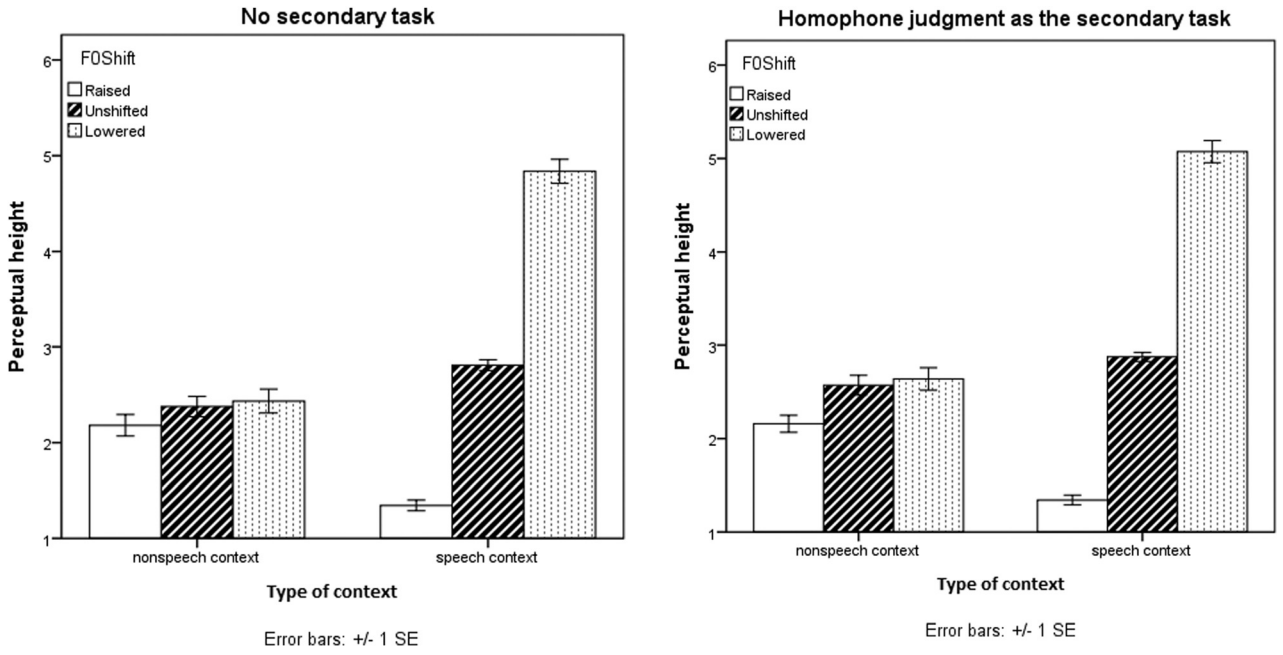

FIG. 3. Average perceptual heights of the word identification task in the single-task paradigm (a) and in the dual-task paradigm with homophone judgment as the secondary task (b).

A four-way repeated-measures analysis of variance (ANOVA), with Greenhouse–Geisser corrections when appropriate, was conducted on perceptual height data with *task paradigm* (single and dual), *context type* (nonspeech and speech), *F0 shift* (raised, unshifted, and lowered), and *talker* (FH, FL, MH, and ML) as within-subjects factors. There were significant main effects of *context type* [F (1, 17) = 62.52, $P < 0.001$], *F0 shift* [F (2, 34) = 245.3, $P < 0.001$], and *talker* [F (3, 51) = 25.8, $P < 0.001$]. The main effect of *task paradigm* was not statistically significant, and it was also not involved in any significant interactions. Meanwhile, participants achieved a high accuracy in the homophone judgment task (90% on average). These results indicate that the secondary task might not affect lexical tone normalization at all.

Besides, there were significant two-way interactions: *context type* by *F0 shift* [F (2, 34) = 180.1; $P < 0.001$], *talker* by *context type* [F (3, 51) = 20.61, $P < 0.001$], and *talker* by *F0 shift* [F (6, 102) = 2.99, $P < 0.05$]. No significant three- or four-way interactions were found. A simple main effect analysis with Bonferroni adjustment was conducted on the interaction of *context type* by *F0 shift*. The results showed that in nonspeech contexts, raising $F0$ caused listeners to demonstrate significantly lower perceptual heights [mean value (M) = 2.17, standard error (SE) = 0.11)] compared to conditions in which $F0$'s were either unshifted ($M = 2.48$, SE = 0.1; $P < 0.001$) or lowered ($M = 2.54$, SE = 0.14; $P < 0.05$). No significance was found between the perceptual heights obtained in the $F0$-unshifted and $F0$-lowered conditions of the nonspeech contexts ($P = 0.414$). In speech contexts, however, listeners' perceptual heights differed significantly across $F0$ conditions (all P's < 0.001), which were 1.34 in the $F0$-raised condition, 2.85 in the $F0$-unshifted condition, and 4.96 in the $F0$-lowered condition.

The simple main effect analysis on the interaction of *talker* by *context* showed that in nonspeech contexts, except MH vs ML ($P = 0.48$), the perceptual height differences of other talker pairs all achieved the significance level (P's < 0.05). However, in speech contexts, the perceptual heights of four talkers were all around 3, indicating that the decisions were largely made by calculating the relative pitch height against the speech contexts, but not by talker' physical $F0$ values.

The simple main effect analysis on the interaction of *talker* by *F0 shift* was shown as below. In the $F0$-unshifted condition, although the physical $F0$ value of FH (246 Hz) is higher that MH (143 Hz), the perceptual heights of FH ($M = 3.07$, SE = 0.11) vs MH ($M = 2.69$, SE = 0.12; $P = 0.11$) were not significantly different. This was also the case for FL (210 Hz; $M = 2.29$, SE = 0.1) vs ML (123 Hz; $M = 2.62$, SE = 0.1; $P = 0.08$). However, the perceptual height of FH was significantly higher than FL ($M = 2.29$, SE = 0.1; $P < 0.05$). It seems that listeners have different expectations toward males' and females' pitch heights and they will judge unfamiliar talkers' pitch heights against the gender-specific expectations (Honorof and Whalen, 2005; Lee, 2009; Bishop and Keating, 2012; Zhang *et al.*, 2012). The perceptual heights of MH and ML were not significantly different ($P = 1$). This might be because their pitch heights (MH: 143 Hz; ML: 123 Hz) were somewhat close to each other.

2. Experiment II: Pitch location judgment task

Figure 4 illustrates the perceptual heights obtained in experiment II. There is a clear contrastive context effect in the nonspeech contexts. Perceptual heights were submitted to four-way repeated-measures ANOVA with *task paradigm*, *context type*, *F0 shift*, and *talker* as within-subjects factors.

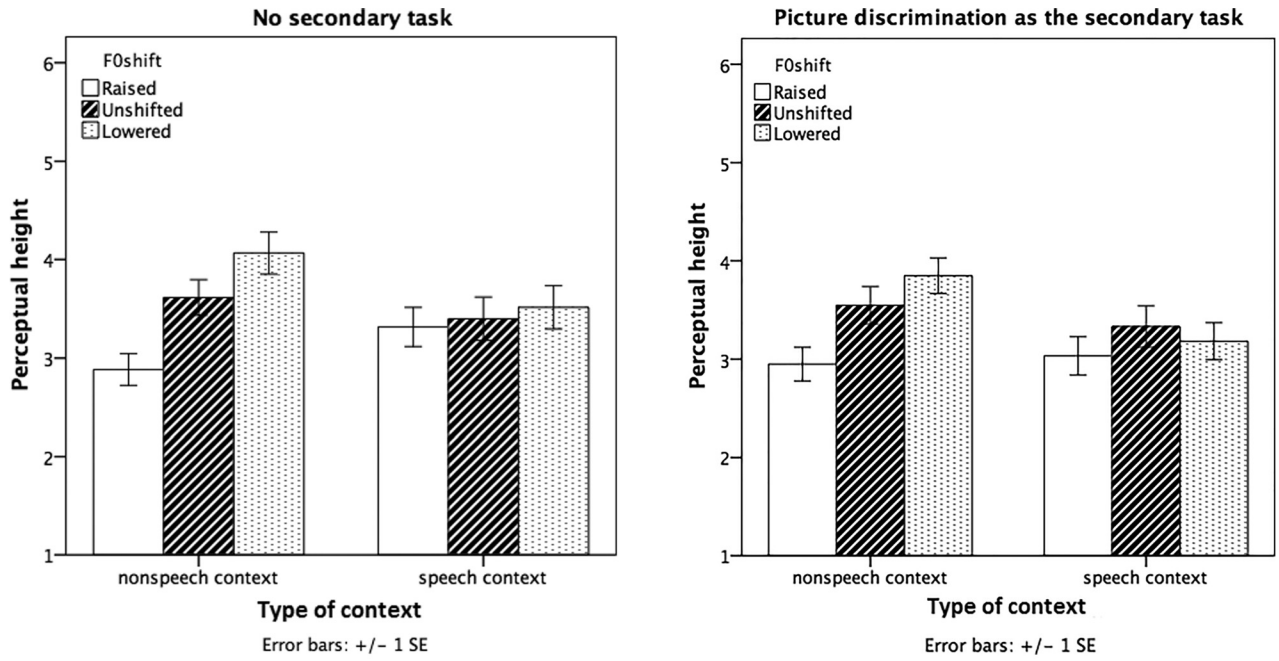

FIG. 4. Average perceptual heights of the pitch location judgment task in the single-task paradigm (a) and in the dual-task paradigm with picture discrimination as the secondary task (b).

The Greenhouse–Geisser method was used to correct violations of sphericity, where appropriate. The analysis revealed significant main effects of *F0 shift* [F (2, 28) = 12.65, $P < 0.001$] and *talker* [F (3, 42) = 19.85; $P < 0.001$]. *Task paradigm* was again not involved in any significant effects, which points to the null effect of focal attention on relative pitch perception. Similar to experiment I, participants achieved high accuracy in the picture discrimination task, with a mean score of 90%.

Additionally, there were significant two-way interactions: *context type* by *F0 shift* [F (2, 28) = 12.53, $P < 0.001$] and *talker* by *context type* [F (3, 42) = 7.73, $P < 0.001$]. No significant three- or four-way interactions were found. According to the simple main effect analysis on the interaction of *context type* by *F0 shift*, *F*0 shifts did not cause the perceptual heights to shift significantly in the speech contexts. However, such manipulations did yield significant differences in the nonspeech contexts (all P's < 0.05), with perceptual heights reaching 2.87 in the *F*0-raised condition, 3.51 in the *F*0-unshifted condition, and 3.91 in the *F*0-lowered condition.

The simple main effect analysis on the interaction of *talker* by *context type* showed that in the speech contexts, except FL ($M = 3.81$, SE = 0.3) vs MH ($M = 2.85$, SE = 0.3; $P = 0.153$), the perceptual height differences of other talker pairs all achieved the significance level (P's < 0.05). In the nonspeech contexts, except FL ($M = 3.62$, SE = 0.26) vs MH ($M = 3.14$, SE = 0.3; $P = 0.45$) and FL vs ML ($M = 2.74$, SE = 0.32; $P = 0.09$), the perceptual height differences of other talker pairs all achieved significance level (P's < 0.05). The significant inter-talker differences suggested that the nonlinguistic pitch normalization is affected more by talkers' physical *F*0 values, but less by the external context cues.

3. Normalization of lexical tones and nonlinguistic pitch contours

Context effect was observed in the normalization of both lexical tones and nonlinguistic pitch contours with proper contexts. To test whether or not there was a significant difference between these two types of normalization (i.e., the normalization of lexical tones and the normalization of nonlinguistic pitch contours), a two-way repeated-measures ANOVA was carried out on participants' perceptual heights in the single-task condition. Only two types of utterances were considered in the analysis: the utterances composed by speech contexts and speech targets, and the utterances composed by nonspeech contexts and nonspeech targets, in which robust context effects emerged. *F0 shift* (raised, unshifted, and lowered) was defined as the within-subject factor and *normalization target* (lexical tones and nonlinguistic pitch contours) was defined as the between-subject factor. Statistical results were summarized in Table I. For every *F*0-shift condition, there was a significant difference in perceptual heights between two types of normalization target. The normalization of lexical tones was much closer to the expected value (i.e., 6 in the lowered, 3 in the unshifted, and 1 in the raised contexts, respectively) compared to those obtained in the non-linguistic pitch normalization.

In summary, both the normalization of lexical tones and nonlinguistic pitch contours showed a contrastive context effect. For lexical tone normalization, while speech contexts consistently improved subjects' tone perception, nonspeech contexts only showed such an effect in the *F*0-raised condition. However, when it came to nonlinguistic pitch contours, it was the nonspeech contexts that drove the significance of the contrastive context effect. The context effect observed

TABLE I. Perceptual height obtained in the lexical tone normalization and nonlinguistic pitch contour normalization (single-task paradigm) tasks. *P*-values were obtained from a two-way repeated-measures ANOVA.

*F*0 shift	Normalization target	Mean	SE	*P*-value
Raised	Lexical tones	1.35	0.19	<0.001
	Nonlinguistic pitch contours	2.75	0.21	
Unshifted	Lexical tones	2.81	0.2	<0.05
	Nonlinguistic pitch contours	3.45	0.21	
Lowered	Lexical tones	4.84	0.26	<0.05
	Nonlinguistic pitch contours	3.88	0.28	

for lexical tones was significantly larger in magnitude than that for nonlinguistic pitch contours in every *F*0-shift condition. Additionally, the physical *F*0 values and the gender of talkers exerted a notable effect on the pitch normalization even when the external context cues are available. Meanwhile, secondary tasks imposed little effect on the extraction of contextual information during both lexical tone and nonlinguistic pitch normalization.

B. Identification rate analysis

Figure 5 illustrates the identification rates participants obtained in the word identification tasks and in the pitch location judgment tasks. Results are shown as functions of contexts and *F*0 shifts. Since the *task paradigm* hardly affected the experimental results (see the analysis below), the identification rates in Fig. 5 were not further divided into two task paradigms. As noted before, the identification rate analysis can reveal the context effect in the *F*0-unshifted condition. The mid-level tone responses were given most frequently for the *F*0-unshifted condition in the word identification task (82% in the speech contexts and 54.3% in the nonspeech contexts), showing the expected context effect. This was also true for the *F*0-unshifted nonspeech contexts in the pitch location judgment task (46.7% mid-level pitch responses). However, the context effect was not obvious for *F*0-unshifted speech contexts in the pitch location judgment task (35.4% mid-level pitch responses).

1. Experiment I: Word identification task

Participants' identification rates were submitted to a four-way repeated-measures ANOVA with *task paradigm* (single and dual), *context type* (speech and nonspeech), *F0 shift* (raised, unshifted, and lowered), and *talker* (FH, FL, MH, and ML) as within-subjects factors. The results reveal significant main effects of *context type* [F (1, 17) = 236.74, $P < 0.001$], *F0 shift* [F (2, 34) = 16.04, $P < 0.001$], and *talker* [F (3, 51) = 3.75, $P < 0.05$].

There were significant two-way interactions: *context type* by *F0 shift* [F (2, 34) = 18.04, $P < 0.001$] and *talker* by *F0 shift* [F (6, 102) = 17.61, $P < 0.001$]. The simple main effect analysis on the interaction of *context type* by *F0 shift* showed that, for every *F*0 shift condition, the identification rate in the speech contexts was significantly higher than in the nonspeech contexts [P's < 0.01, see Fig. 5(a)], suggesting that lexical tone normalization preferred the speech contexts.

The simple mean effect analysis was also conducted for the interaction of *talker* by *F0 shift*. In the *F*0-unshifted condition, FH ($M = 0.8$, SE = 0.04) was significantly better than another three talkers (P's < 0.05). On the contrary, FL ($M = 0.56$, SE = 0.04) was the worst (P's < 0.05). MH ($M = 0.68$, SE = 0.04) and ML ($M = 0.69$, SE = 0.04) were not significantly different from each other ($P = 1$). Talkers whose pitch ranges are closer to the population mean are

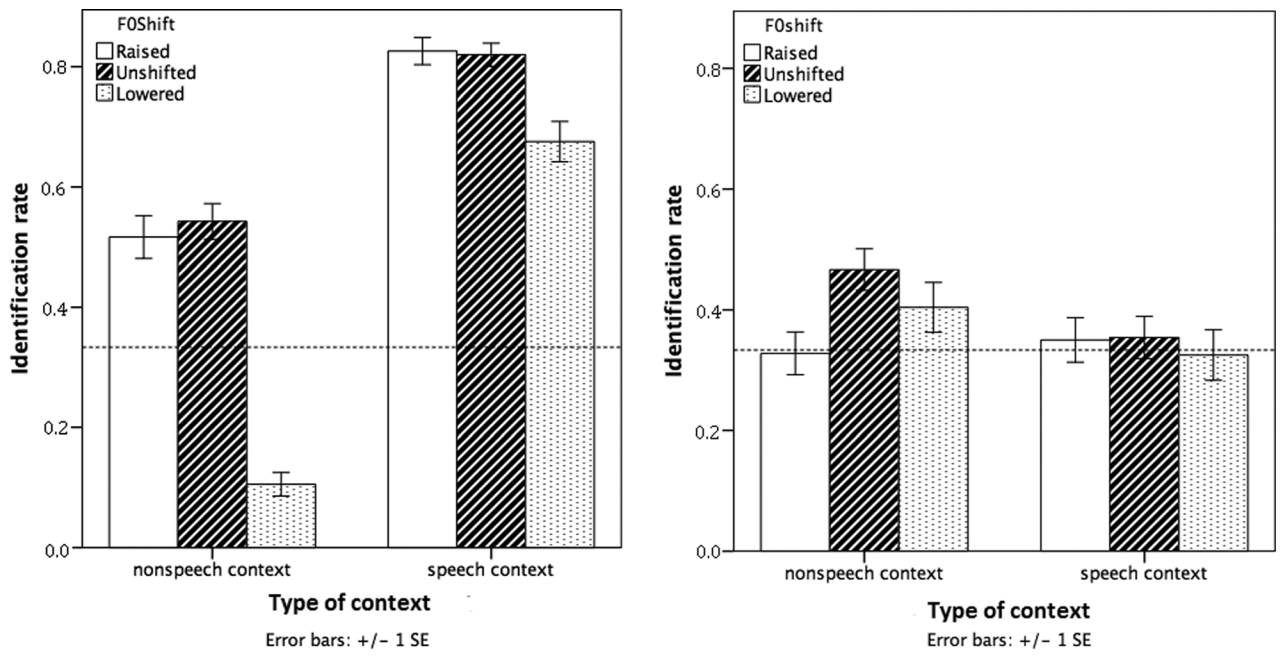

FIG. 5. Average identification rates in the word identification task (a) and in the pitch location judgment task (b). The dashed lines in the two graphs refer to the chance level (33%).

comparatively easier to recognize (Zhang *et al.*, 2012). The comparison between the average *F*0 range of Cantonese speakers (female: 200–290 Hz, male: 110–160 Hz from Peng *et al.*, 2012) and the pitch ranges of the informants used in the present study (FH: 198–294 Hz, FL: 166–279 Hz, MH: 112–194 Hz, and ML: 96–151 Hz) shows that FH is closet to the average pitch range, FL is relatively far away from the average, and the two male informants are somewhat in between. This might explain why FH achieved the highest accuracy and FL is the most difficult to be identified. In the *F*0-lowered and the *F*0-raised conditions, except FH vs MH, FL vs ML, and MH vs ML, the identification rates of other talker pairs all achieved the significance level (P's < 0.05).

There was also a three-way interaction, *talker* by *F0 shift* by *context* [F (6, 102) $= 10.09$, $P < 0.001$]. The results in speech contexts were reported first. In most cases, the identification rates of four talkers were comparatively high (all around 80%) and did not show a significant difference, suggesting that the speech contexts were effective in normalizing the between-talker variations. However, in the *F*0-unshifted condition, FH ($M = 0.88$, SE $= 0.03$) was significantly better than FL ($M = 0.78$, SE $= 0.04$) ($P < 0.05$). A further analysis revealed that FL in the *F*0-unshifted condition was most frequently misidentified as the low level tone (18%). Besides, FH ($M = 0.78$, SE $= 0.06$) was also significantly better than both FL ($M = 0.58$, SE $= 0.08$) ($P < 0.05$) and ML ($M = 0.64$, SE $= 0.07$; $P < 0.05$) in the *F*0-lowered condition. The target /ji33/ was expected to be perceived as the high level tone in the *F*0-lowered condition. The target stimuli produced by FH whose average *F*0 height is highest should be comparatively easier to be perceived as the high level tone. Such perceptual results indicate that subjects' tone identification was somewhat affected by the physical *F*0 values of the targets, and that the intrinsic cues like pitch height of the targets and the extrinsic cues (the contextual information) interact with each other in pitch normalization. The following are the results in nonspeech contexts. In the *F*0-unshifted condition, except MH vs ML ($P = 1$), the identification rates of other talker pairs all achieved the significance level (P's < 0.05). In the *F*0-lowered condition, the identification rates of four talkers were all lower than the chance level 33%. In the *F*0-rasied condition, except FH vs MH ($P = 0.1$) and MH vs ML ($P = 0.32$), the identification rates of other talker pairs all achieved the significance level (P's < 0.05). As the results showed, talker normalization may be harder in nonspeech contexts.

2. Experiment II: Pitch location judgment task

A four-way repeated-measures ANOVA was also conducted for the identification rates obtained in the pitch location judgment task, with *task paradigm*, *context type*, *F0 shift*, and *talker* as within-subjects factors. The results reveal a significant main effect of *context type* [F (1, 14) $= 13.34$, $P < 0.01$].

There were significant two-way interactions: *talker* by *F0 shift* [F (6, 84) $= 7.07$, $P < 0.01$] and *task* by *F0 shift* [F (2, 28) $= 8.24$, $P < 0.05$]. A simple main effect analysis on the interaction, *talker* by *F0 shift*, showed that talkers were not significantly different from each other in the *F*0-unshifted condition. In the *F*0-lowered condition, FH is the easiest to be recognized ($M = 0.54$, SE $= 0.08$). In the *F*0-raised condition, ML became the easiest one ($M = 0.55$, SE $= 0.07$). The perceptual results suggested that talkers' physical *F*0 values also affected the nonlinguistic pitch normalization.

The simple main effect analysis on the interaction of *task* by *F0 shift* revealed that in the *F*0-unshifted condition, the accuracies in two task paradigms were not significantly different from each other ($P = 0.97$). In the *F*0-rasied condition, the accuracy in the dual-task paradigm ($M = 0.38$, SE $= 0.04$) was significantly better than that in the single-task paradigm ($M = 0.3$, SE $= 0.04$; $P < 0.05$). However, in the *F*0-lowered condition the accuracy in the dual-task paradigm ($M = 0.31$, SE $= 0.05$) became significantly worse than that in the single-task paradigm ($M = 0.42$, SE $= 0.07$; $P < 0.05$). Considering that the task paradigm did not show consistent effects on pitch perception and that it was not involved in any other significant effects, this occasional significant two-way interaction might not be reliable.

The two-way interaction of *context type* by *F0 shift* was marginally significant [F (2, 28) $= 3.52$, $P = 0.051$]. The simple main effect analysis showed that, for the *F*0-lowered and *F*0-unshifted conditions, the identification rates of the nonlinguistic pitches were significantly higher when the contexts were nonspeech [P's < 0.01, see Fig. 5(b)]. However, the speech and nonspeech contexts contributed equally to the nonlinguistic pitch normalization in the *F*0-raised condition.

The identification rate analysis further revealed that lexical tone normalization preferred the speech contexts, while the nonspeech contexts were more helpful in the normalization of nonlinguistic pitch contours. Besides, consistent with Zhang *et al.* (2012), talkers' relative pitch heights are easier to recognize if their pitch ranges are closer to the population mean.

IV. DISCUSSION

A. The potential cause of the speech-specific context effect on lexical tone normalization

Rather than merely replicating the speech-specific context effect on lexical tone normalization (Francis *et al.*, 2006; Zhang *et al.*, 2012), the results of the present study also provide clues to the probable cause of such an effect. As previously mentioned, there are at least three possible factors that may contribute to this effect. Francis *et al.* (2006) posited that listeners might have selectively ignored nonspeech contexts. Without focal attention, the pitch information in the nonspeech contexts cannot be properly processed. To test this assumption, the present study manipulated the focal attention in both speech and nonspeech contexts. In order to prevent subjects from actively perceiving the simultaneously-played auditory contextual stimuli, the secondary tasks, especially the homophone judgment task, were made highly attention-intensive. First of all, the Chinese characters used in each trial of the homophone judgment task were different so that subjects have to pay attention to each trial. Besides, considering that phonological information is

activated in the N400 time window for the frequently-used Chinese characters in the Chinese homophone judgment task (Zhang *et al.*, 2009), and that a comparative short duration of 300 ms was adopted to present the Chinese characters, subjects could hardly perceive the visual characters clearly and retrieve the phonological information successfully without focal attention. The high accuracies in both the homophone judgment task and the picture discrimination task (all above 90%) suggested that participants did fully attend to the secondary tasks as instructed and the manipulation of the attention on the contextual perception was effective. However, even in the attention-deprived condition, speech contexts still strongly influenced lexical tone perception [see Fig. 3(b)], suggesting that at least the context processing in the pitch normalization does not heavily rely on focal attention.

Furthermore, Lee *et al.* (1996) believed that listeners' rich experiences with speech gave speech perception a marked advantage over nonspeech perception. However, as shown by the results of the pitch location judgment task, the familiarity does not always play a positive role. The context effect could instead be stronger in the nonspeech condition, if the target was likewise a nonlinguistic stimulus (see Fig. 4). In other words, nonspeech contexts may be the preferred medium for perceptual normalization in the appropriate conditions, and prior perceptual experience is not a prerequisite of effective pitch normalization.

In summary, the two perceptual experiments in the present study revealed the following results: (1) The speech-specific context effect surfaced even when subjects' focal attention was deprived during the presentation of contexts; (2) when identifying nonlinguistic pitch contours, listeners relied more heavily on nonspeech contexts, rather than on speech contexts with which they had more experience. These results suggest that focal attention (Francis *et al.*, 2006) and the degree of familiarity (Lee *et al.*, 1996) may exert less effect on the unequal effect of speech and nonspeech contexts compared with the speech-specific mechanism. The evidence for the speech-specific mechanism mainly lies in two aspects. First, the context effect is more conspicuous on speech perception than that on nonspeech perception. For every *F0 shift*, the normalization of lexical tones was much closer to the expected pitch height, compared to the results obtained in the non-linguistic pitch normalization task (see Table I). Second, speech perception and nonspeech perception benefit differentially from linguistic and nonlinguistic contexts. Specifically, participants relied more heavily on speech contexts to identify the relative pitch heights of speech targets. Interestingly, nonspeech analogs became the most effective contexts when it came to the nonlinguistic pitch perception. These results suggest that speech and nonspeech may be processed at least partly by different mechanisms and that the superior context effect of speech is mostly caused by the speech-specific mechanism.

B. The congruency effect in normalizing lexical tones and nonlinguistic pitches

By testing the normalization of linguistic and nonlinguistic pitch contours in speech and nonspeech contexts, the present study reveals a strong *congruency effect*: The pitch height of the target sound is easier to recognize when the context and the target are of the same nature. As can be seen from the lexical tone normalization task, *F*0 shifts can be distinguished easily in congruent situations (speech targets paired with speech contexts). Incongruent utterances (speech targets paired with nonspeech contexts), however, posed difficulties to tone normalization. Such a congruency effect also applies to nonlinguistic pitch judgment. A statistically significant context effect was elicited by the utterances composed by nonspeech targets and nonspeech contexts but not in the incongruent condition (nonspeech targets paired with speech contexts).

The speech-specific context effect (Zhang *et al.*, 2012), which predicts that only speech contexts can significantly improve lexical tone perception, might be a reflection of the congruency effect in processing speech sound. Results of the present study further demonstrate that this congruency effect also applies to nonspeech processing: The identification of nonlinguistic pitch contours is pronouncedly facilitated by nonspeech contexts but not by speech contexts. This is consistent with the postulation that speech sounds and nonspeech sounds are partly processed by different mechanisms and that the information processed by the same mechanism can be integrated more easily. The fact that a slight contextual effect also appeared in the incongruent conditions shows that information exchange across two mechanisms is possible but limited.

There is also another potential congruency [i.e., a match between the nature of the task (linguistic or nonlinguistic) and the context (speech or nonspeech)] that may affect the pitch normalization. As the results suggested, speech contexts were more helpful in the word identification task (a linguistic task), whereas the nonlinguistic task (i.e., the pitch judgment task) preferred the nonspeech contexts. It seems that the perceptual results may be mediated by the nature of the tasks in a top-down manner (Zekveld *et al.*, 2006). Further studies need to be carried out to clarify how and to what extent the nature of the tasks affects the pitch normalization.

C. The automaticity of perceiving contextual cues in pitch normalization

Human information processing can be divided into controlled processing and automatic processing (Schneider and Shiffrin, 1977). Automatic processing requires relatively little attention and few processing resources. Its operation will not be interfered by other concurrent information processing. By contrast, controlled processing is intentional and is constrained by attentional resources. As stated at the beginning of this session, the secondary tasks in the dual-task blocks were designed to be highly attention-intensive and were expected to compete for attentional resources used for perceiving contextual cues. However, in the attention-deprived condition (i.e., in the dual-task condition), the normalization results of lexical tones and nonlinguistic pitches decreased only slightly. This demonstrates that one crucial step in speech normalization, the extraction of contextual information, does not rely heavily on attentional and cognitive resources and is likely an instance of automatic processing.

However, the entire process of extrinsic speech normalization might be controlled. The identification of the vowels embedded in a "p_p" structure was more accurate in single-talker conditions than in mixed-talker conditions (Verbrugge *et al.*, 1976). The normalization of consonants in CV syllables was also subject to variations of attentional resources and might be processed in a controlled manner (Nusbaum and Morin, 1992). It seems that if the mapping between speech signals and their mental representations is stable (e.g., when there is no change in talker identity), signals may be recognized automatically; otherwise, a controlled process is triggered to acquire talker-specific acoustic properties (Nusbaum and Schwab, 1986; Nusbaum and Morin, 1992).

Hence, these results together might suggest that, while the process of speech normalization requires focal attention, the extraction of useful contextual information, one of the steps in perceptual normalization, may be automatic. Another possibility could be that human brains process the segmental and the suprasegmental components differently, since Nusbaum and Morin (1992) studied the perceptual normalization of segments, while the present study concentrated on the suprasegmental level. Further studies need to be carried out to specify which phases during perceptual normalization are controlled processes and whether or not focal attention plays different roles in normalizing segmental and suprasegmental components.

V. CONCLUSION

Previous studies have proposed some potential factors that may contribute to the speech-specific context effect on lexical tone normalization. By testing lexical tone and nonlinguistic pitch normalization under different contextual and attentional conditions, the present study shows that the speech-specific mechanism is the most likely factor. Moreover, the current study extends previous findings by showing that the congruency effect found in the lexical tone perception likewise applies to the normalization of nonlinguistic pitch contours, suggesting that contexts of the same nature as targets can be a better reference for normalizing pitches of the target sounds. The congruency effect also suggests that linguistic and nonlinguistic pitch processing may partly depend on distinct neural mechanisms and that information processed by the same mechanism can be integrated more efficiently than that processed by different mechanisms. Although the integration of information across mechanisms was also observed in this study, its magnitude was much smaller than the integration carried out by a single mechanism. Finally, our findings indicate that the relevant contextual information for pitch normalizations is likely processed automatically, regardless of the types of contexts in question.

ACKNOWLEDGMENTS

This study was supported in part by grants from the National Natural Science Foundation of China (NSFC: Grant Nos. 61135003 and 11474300), and Research Grant Council of Hong Kong (GRF: Grant Nos. 448413 and 14408914).

Bishop, J., and Keating, P. (**2012**). "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," J. Acoust. Soc. Am. **132**, 1100–1112.
Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer (Version 5.3.23) [Computer program]," http://www.praat.org (Last viewed August 7, 2012).
Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA), pp. 529–594.
Chao, Y. R. (**1969**). *Cantonese Primer* (Greenwood Press, New York), pp. 1–242.
Diehl, R. L., and Kluender, K. R. (**1989**). "On the objects of speech perception," Ecological Psychol. **1**, 121–144.
Fedorenko, E., Behr, M. K., and Kanwisher, N. (**2011**). "Functional specificity for high-level linguistic processing in the human brain," Proc. Natl. Acad. Sci. U.S.A. **108**, 16428–16433.
Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., and Chu, P. C. Y. (**2006**). "Extrinsic context affects perceptual normalization of lexical tone," J. Acoust. Soc. Am. **119**, 1712–1726.
Garrett, K. L., and Healey, E. C. (**1987**). "An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day," J. Acoust. Soc. Am. **82**, 58–62.
Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **16**, 78–80.
Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**, 2193–2200.
Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**, 3983–3994.
Huang, J., and Holt, L. L. (**2011**). "Evidence for the central origin of lexical tone normalization," J. Acoust. Soc. Am. **129**, 1145–1148.
Hugdahl, K., Thomsen, T., Ersland, L., Rimol, L. M., and Niemic, J. (**2003**). "The effects of attention on speech perception: An fMRI study," Brain Lang. **85**, 37–48.
Johnson, K., and Mullennix, J. W. (eds.) (**1997**). *Talker Variability in Speech Processing* (Academic Press, San Diego, CA), pp. 1–237.
Joos, M. A. (**1948**). "Acoustic phonetics," Language **24**(2), 1–136.
Kaan, E., Wayland, R., Bao, M., and Barkley, C. M. (**2007**). "Effects of native language and training on lexical tone perception: An event-related potential study," Brain Res. **1148**, 113–122.
Leather, J. (**1983**). "Speaker normalization in perception of lexical tone," J. Phonetics **11**, 373–382.
Lee, C.-Y. (**2009**). "Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study," J. Acoust. Soc. Am. **125**, 1125–1137.
Lee, Y. S., Vakoch, D. A., and Wurm, L. H. (**1996**). "Tone perception in Cantonese and Mandarin: A cross-linguistic comparison," J. Psychol. Res. **25**, 527–542.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.
Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.
Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.
Nusbaum, H. C., and Morin, T. M. (**1992**). "Paying attention to differences among talkers," in *Speech Perception, Speech Production, and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (IOS Press, Amsterdam), pp. 113–134.
Nusbaum, H. C., and Schwab, E. C. (**1986**). "The role of attention and active processing in speech perception," in *Pattern Recognition by Humans and Machines: Speech Perception*, Vol. 1, edited by E. C. Schwab and H. C. Nusbaum (Academic Press, San Diego, CA), pp. 113–157.
Peng, G., Zhang, C. C., Zheng, H. Y., Minett, J. W., and Wang, W. S. Y. (**2012**). "The effect of intertalker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems," J. Speech Lang. Hear. Res. **55**, 579–595.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pike, K. L. (**1948**). *Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion* (University of Michigan Press, Ann Arbor, MI), pp. 3–39.
Protopapas, A., and Lieberman, P. (**1997**). "Fundamental frequency of phonation and perceived emotional stress," J. Acoust. Soc. Am. **101**, 2267–2277.
Schneider, W., and Shiffrin, R. M. (**1977**). "Controlled and automatic human information processing: I. Detection, search, and attention," Psychol. Rev. **84**, 1–66.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **60**, 198–212.
Wang, W. S. Y. (**1972**). "The many uses of F0," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, edited by A. Valdman (Mouton, The Hague), pp. 487–503.
Wang, Y., Jongman, A., and Sereno, J. A. (**2003**). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. **113**, 1033–1043.
Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (**1999**). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. **106**, 3649–3658.
Wayland, R. P., and Guion, S. G. (**2004**). "Training English and Chinese listeners to perceive Thai tones: A preliminary report," Lang. Learn. **54**, 681–712.
Whalen, D. H., Benson, R. R., Richardson, M., Swainson, B., Clark, V. P., Lai, S., Mencl, W. E., Fulbright, R. K., Constable, R. T., and Liberman, A. M. (**2006**). "Differentiation of speech and nonspeech processing within primary auditory cortex," J. Acoust. Soc. Am. **119**, 575–581.
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.
Yip, M. (**2002**). *Tone* (Cambridge University Press, Cambridge), pp. 17–208.
Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., and Schoonhoven, R. (**2006**). "Top-down and bottom-up processes in speech comprehension," NeuroImage **32**, 1826–1836.
Zhang, C. C., Peng, G., and Wang, W. S. Y. (**2012**). "Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones," J. Acoust. Soc. Am. **132**, 1088–1099.
Zhang, C. C., Peng, G., and Wang, W. S. Y. (**2013**). "Achieving constancy in spoken word identification: Time course of talker normalization," Brain Lang. **126**, 193–202.
Zhang, Q., Zhang, J. X., and Kong, L. Y. (**2009**). "An ERP study on the time course of phonological and semantic activation in Chinese word recognition," Int. J. Psychophys. **73**, 235–245.