Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones

Caicai Zhang,[a] Gang Peng,[b] and William S-Y. Wang
Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong Special Administrative Region

(Received 11 October 2011; revised 12 March 2012; accepted 12 June 2012)

Context is important for recovering language information from talker-induced variability in acoustic signals. In tone perception, previous studies reported similar effects of speech and nonspeech contexts in Mandarin, supporting a general perceptual mechanism underlying tone normalization. However, no supportive evidence was obtained in Cantonese, also a tone language. Moreover, no study has compared speech and nonspeech contexts in the multi-talker condition, which is essential for exploring the normalization mechanism of inter-talker variability in speaking F0. The other question is whether a talker's full F0 range and mean F0 equally facilitate normalization. To answer these questions, this study examines the effects of four context conditions (speech/nonspeech × F0 contour/mean F0) in the multi-talker condition in Cantonese. Results show that raising and lowering the F0 of speech contexts change the perception of identical stimuli from mid level tone to low and high level tone, whereas nonspeech contexts only mildly increase the identification preference. It supports the speech-specific mechanism of tone normalization. Moreover, speech context with flattened F0 trajectory, which neutralizes cues of a talker's full F0 range, fails to facilitate normalization in some conditions, implying that a talker's mean F0 is less efficient for minimizing talker-induced lexical ambiguity in tone perception.
[http://dx.doi.org/10.1121/1.4731470]

PACS number(s): 43.71.Bp, 43.71.Es, 43.71.An, 43.71.Sy [MAH] Pages: 1088–1099

I. INTRODUCTION

There is great inter- and intra-talker variation in speech production. The same word uttered by different talkers (inter-talker variability) or by a talker on different occasions (intra-talker variability) shows great acoustic variation. Despite the acoustic variation, listeners can recognize the intended word without much difficulty. This phenomenon is known as talker normalization (Johnson and Mullennix, 1997; Johnson, 2005).

In tone languages, fundamental frequency (F0) is used to distinguish lexical meanings (e.g., Wang, 1972). Due to physiological differences in the vocal apparatus, talkers in a speech community have different speaking F0 (e.g., Rose, 1996). Speaking F0 also varies within the same talker across the day (Garrett and Healey, 1987) and changes as a function of emotional mood (Protopapas and Lieberman, 1997) and other factors. Inter- and intra-talker difference in speaking F0 gives rise to varied F0 realizations of a tone. In this study, talker normalization is investigated in tone perception, i.e., how listeners recognize the same tone produced by different talkers. This process is referred to as "tone normalization" hereafter (Francis *et al.*, 2006).

Rich level tones in Cantonese provide an optimal window to probe the mechanism of tone normalization. There are three level tones in Cantonese, high level tone (e.g., /ji55/ 醫 "a doctor"), mid level tone (e.g., /ji33/ 意 "meaning"), and low level tone (e.g., /ji22/ 二 "two") (Bauer and Benedict, 1997). Inter- and intra-talker variation causes overlapping in the acoustic realization of Cantonese level tones (Rose, 1996), leading to ambiguity in perception (Wong and Diehl, 2003; Peng *et al.*, 2012). To resolve the perceptual ambiguity of words carrying level tones, the context with cues of a talker's speaking F0 is essential (Wong and Diehl, 2003; Francis *et al.*, 2006). This study aims to investigate the effect of different types of contexts on the perceptual normalization of inter- and intra-talker variation via Cantonese level tones, which may shed light on the mechanism of tone normalization at large.

A. Phonetic context effect

There are two sources of cues for tone perception, word-internal cues and word-external cues (i.e., contextual cues). Word-internal cues include F0, duration, intensity, voice quality, and other acoustic cues of a word, which all provide information about the tone category, but F0 is the most important correlate for tone perception (Wang, 1972). Often, tone perception is not only based on word-internal F0 cues but also with reference to acoustic cues in the context (Wong and Diehl, 2003; Francis *et al.*, 2006; Huang and Holt, 2009, 2011). The effect of contextual cues on the perception of a target word, a phenomenon known as the phonetic context effect, has been widely investigated in the literature (e.g., Ladefoged and Broadbent, 1957; Lin and Wang, 1984; Repp, 1982). In tone normalization, the context with cues of a talker's speaking F0 also plays an important role.

[a]Author to whom correspondence should be addressed. Electronic mail: yzcelia@gmail.com
[b]Also at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China.

Lin and Wang (1984) have investigated how the relative F0 height of the context affected the perception of Mandarin tones. An identical target word attached to a word produced with either high or low F0 was perceived as having different tones. These authors found a contrastive context effect, i.e., more low tone responses were elicited in the high F0 context, whereas more high tone responses were elicited in the low F0 context. The contrastive context effect demonstrated that listeners rescaled the pitch percept of a target word relative to the pitch of its immediate context. Findings of this study shed light on the normalization of intra-talker variability. As mentioned earlier, there is variation in speaking F0 within a talker (Garrett and Healey, 1987; Protopapas and Lieberman, 1997). Listeners may evaluate a talker's speaking F0 from the immediate context, which reflected the most up-to-date status of the dynamic variation of a talker's speaking F0. For example, the low F0 context implied that this talker spoke with a low F0 at a particular time, which then led listeners to judge a target word as carrying a high tone.

Leather (1983) explicitly examined the effect of contextual cues on the normalization of inter-talker variability. Talker-ambiguous Mandarin tone stimuli were embedded in the speech utterances produced by two male talkers with different F0 ranges. Identical stimuli were perceived as different tones, depending on which talker was perceived to "produce" them. Leather's findings provided initial evidence for the effect of the context on talker normalization.

Moore and Jongman (1997) found that the phonetic context with F0 information of a specific talker affected the identification of talker-ambiguous tone stimuli in a contrastive way, i.e., the same stimulus was identified as having a low tone if the context was produced by the talker with high mean F0 and as having a high tone if the context was produced by a talker with a low F0.

Lee *et al.* (2009) reported that the effect of phonetic context was modulated by the language background of listeners. Native listeners facilitated more from the phonetic context than non-native listeners in tone identification in the multi-talker condition.

The above-noted studies examined tone normalization in Mandarin. Wong and Diehl (2003) and Francis *et al.* (2006) investigated this issue in Cantonese. These studies also reported a contrastive context effect, i.e., raising or lowering the F0 of the context shifted the perception of identical target words in a contrastive way. Moreover, within the carrier sentence, words adjacent to the target word had a greater effect than those far away (Wong and Diehl, 2003). The effects of the contexts preceding and following the target word were also different. When the F0 of the preceding and following contexts was shifted in different directions, giving contradictory information, listeners tended to rely on the following context for tone normalization (Francis *et al.*, 2006).

In summary, most previous studies found a contrastive context effect in tone perception (except for Fox and Qi, 1990, who reported a marginally assimilatory effect), suggesting that the contextual F0 is used as a reference for perceiving the tone of the target word.

B. Nonspeech context

Consistent findings about the effect of speech context raise a question regarding nonspeech context—does nonspeech context have a similar effect on tone normalization? There are two different views about the processing of speech and nonspeech sounds in the literature. One view holds that only speech sounds are processed via the phonetic module (e.g. Liberman *et al.*, 1967; Liberman and Mattingly, 1985), different from nonspeech sounds. The other view proposes a general perceptual mechanism that underlies the processing of both speech and nonspeech sounds (e.g., Holt, 2006a,b; Holt and Lotto, 2008). It is yet unclear whether the processing of suprasegmental properties of speech such as lexical tones recruits the speech-specific mechanism (Francis *et al.*, 2006), or whether the normalization of lexical tones is mediated by the general perceptual mechanism (Huang and Holt, 2009, 2011).

Huang and Holt (2009, 2011) embedded a continuum of eight-step tone stimuli (ranging from Mandarin Tone 1 to Tone 2) in speech and nonspeech contexts. These authors found that nonspeech contexts (harmonic tones, and pure tone contexts) that modeled the mean F0 of speech contexts elicited a qualitatively similar, although quantitatively reduced effect on the identification of Mandarin tone stimuli. Similar effects of speech and nonspeech contexts led the authors to suggest that the general perceptual mechanism is engaged in lexical tone perception.

Francis *et al.* (2006) found that the F0 trajectory extracted from the speech context and superimposed on a [ə] sound generated with the "hummed" neutral vocal tract had no effect on the perception of Cantonese level tones, unlike the speech context. Despite some uncertainty about the interpretation of the [ə] sound, i.e., whether it was a nonspeech sound or a nonnative speech sound for Cantonese listeners, this study casts doubt on an exclusive account of the general perceptual mechanism.

How to explain the contradictory findings in previous studies? One difference between Huang and Holt (2009, 2011) and Francis *et al.* (2006) lies in the tone language that was examined (Mandarin vs Cantonese). However, it is difficult to attribute the contradictory findings simply to language difference. Peng *et al.* (2012) found that Mandarin listeners but not Cantonese listeners showed stable talker normalization without contextual cues, implying that the context is especially important for resolving the ambiguity of Cantonese tones (Wong and Diehl, 2003; Francis *et al.*, 2006). If language difference indeed plays a role, it is reasonable to expect that Cantonese listeners would show a stronger context effect than Mandarin listeners.

The previous studies also differed in the type of nonspeech contexts used, but this factor cannot explain the discrepancy either. Huang and Holt (2009) used harmonic tones and pure tone contexts, which were more nonspeech-like than the [ə] context (ambiguous between speech and nonspeech sounds) used by Francis *et al.* (2006). If the type of nonspeech contexts influences tone perception, the effect of nonspeech contexts should also be found in Francis *et al.* (2006).

In Huang and Holt (2009), the effect of nonspeech contexts was analyzed as how the relative F0 height of nonspeech contexts (high or low F0) modulated the identification of a continuum of tone stimuli. The effect of nonspeech contexts was greatest for tone stimuli ambiguous between Tone 1 and Tone 2 (i.e., stimuli in the middle of a continuum), where the high F0 context mildly increased the ratio of Tone 2 responses compared to the low F0 context (e.g., from roughly 50% to 60%). There was barely any effect for unambiguous tone stimuli (i.e., stimuli close to the two ends of a continuum). Francis *et al.* (2006) used a rather different method to analyze the effect of nonspeech contexts. Each tone response was scored (a low level tone response scored as −1, a mid level tone response as 0, and a high level tone response as +1) and the average score of all the responses was calculated for each F0 shift condition (raised, unshifted, and lowered). In other words, this analysis took into account the ratio of all three tone responses in a condition (i.e., not only the ratio of an expected tone response). Although the authors reported no significant change in the average response scores across the F0 shift conditions, it is likely that the F0 shift of nonspeech context also mildly increased the identification ratio of the expected tone response, which somehow failed to surface due to the averaging of the scores of all three tone responses.

To reach the general conclusion that tone normalization is mediated by the general perceptual mechanism, it is crucial to obtain evidence for nonspeech context effect not only in one language. However, so far no supportive evidence is obtained in languages other than Mandarin. The present study aims to reexamine the effects of speech and nonspeech contexts in Cantonese. Moreover, due to the different analysis methods adopted in the previous studies, it is difficult to directly compare the effect of nonspeech contexts. In the present study, the identification rate of tone stimuli is analyzed in order to be comparable with previous studies (Huang and Holt, 2009, 2011).

More importantly, no previous study has compared the effects of speech and nonspeech contexts in the multi-talker condition. At the center of talker normalization lies the issue of inter-talker variability. Although the findings based on one talker are suggestive, it is necessary to take into account the fact that talkers in a speech community have different speaking F0. To obtain substantial evidence for the nature of tone normalization (i.e., speech-specific or general perceptual), it is important to test whether speech and nonspeech contexts equally contribute to the normalization of multiple talkers with different speaking F0.

C. F0 range and mean F0

As reviewed earlier, F0 cues in the context serve as a reference for tone normalization. However, it is less clear what kind of F0 cues contributes to tone normalization. A talker's F0 range computed from the F0 contour in the context (e.g., rising and falling F0 trajectory) may be used as a reference for evaluating the tone of a word (Wong, 1998; Wong and Diehl, 2003). But alternatively, a talker's mean F0 (without explicit information of the upper and lower bounds of F0 range) may be used to normalize talker difference (Huang and Holt, 2009).

Francis *et al.* (2006) investigated the effects of F0 range and mean F0 cues in speech contexts. In the mean F0 condition, the original F0 of a speech sentence was replaced with flattened mean F0, thereby neutralizing cues of the upper and lower bounds of a talker's F0 range. But listeners were able to perceive that this talker had relatively high or low mean F0. Interestingly, the mean F0 condition was even more efficient in eliciting the expected tone responses than the original F0 condition in the perception of Cantonese level tones.

Huang and Holt (2009) synthesized the nonspeech context with a uniform level F0 equal to the mean F0 of speech utterances. As mentioned earlier, the identification of tone stimuli was shifted contrastively according to the F0 of the nonspeech context (high or low F0). The effect of the nonspeech context was qualitatively similar but reduced compared to the speech context.

To summarize, previous studies found that listeners rely on mean F0 cues for tone normalization. But no previous study compared the mean F0 and F0 contour conditions in nonspeech contexts. Nonspeech contexts synthesized with the original F0 contour bears more resemblance to the speech utterances than nonspeech contexts with mean F0. Indeed, Huang and Holt (2009) found that nonspeech contexts with mean F0 showed reduced effect compared to the speech contexts. If the general perceptual mechanism holds, it is likely that speech and nonspeech contexts with the original F0 contour would be similarly efficient in tone normalization.

D. Research aims

This study aims to investigate how different types of contexts affect the perceptual normalization of inter- and intra-talker variation in Cantonese level tones. Two questions are asked:

(1) Is tone normalization mediated by a speech-specific process or a general perceptual process?
(2) Is tone normalization based on a talker's full F0 range computed from the F0 contour in the context, or a talker's mean speaking F0?

This study adopts a within-subjects 2×2 factorial design, *speech type* (speech and nonspeech contexts) × *F0 type* (contexts with F0 contour and flattened mean F0), giving rise to four context conditions. Four native Cantonese speakers (2 female, 2 male) with different F0 ranges are recruited. The preceding two questions are examined by comparing the effects of these four contexts on tone normalization in the multi-talker condition.

For question (1), finding similar effects of speech and nonspeech contexts for all four talkers provides more conclusive evidence for the general perceptual account of speech perception (e.g., Huang and Holt, 2009, 2011). But if the effect is limited to the speech context, it implies that tone normalization is mediated by the speech-specific mechanism (e.g. Francis *et al.*, 2006).

For question (2), if contexts with F0 contour and flattened mean F0 show similar effects, it suggests that a talker's mean F0 suffices for cuing tone normalization (Huang and Holt, 2009; Francis *et al.*, 2006). But if only the context with the F0 contour has an effect, it suggests that a talker's F0 range is essential for tone normalization (e.g., Wong, 1998).

II. METHOD

A. Participants

Sixteen native speakers of Hong Kong Cantonese (8 female, 8 male; mean age = 22.6 yr, SD = 2.6) were paid for their participation in the experiment. No subjects reported hearing impairment or long-term music training. All subjects gave informed consent in compliance with a protocol approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong.

B. Stimuli

Four native Cantonese speakers (2 female, 2 male), all in their early 20 s were recruited to record the speech utterances. Among these four talkers, two talkers within each gender had different but overlapping F0 ranges (female high: 180 ~ 350 Hz, female low: 180 ~ 280 Hz, male high: 110 ~ 190 Hz, male low: 80 ~ 130 Hz).[1] Previous studies showed that F0 range difference between talkers gave rise to perceptual ambiguity without contextual cues (e.g., Peng *et al.*, 2012; Wong and Diehl, 2003). Talker variability in F0 range enables us to examine how efficiently each context resolves the talker-induced lexical ambiguity.

Each talker was asked to read aloud a Cantonese sentence for six times, i.e., 請你讀意嚟聽下 /ts^{h}iŋ25 lei21 tuk22 ji33 lei21 t^{h}iŋ55 ha23/ "please read /ji33/ for me." 意 /ji33/ "meaning" (mid level tone) in the middle of this sentence was the target word, and the remaining part served as the context. The context contained cues of a talker's full F0 range, i.e., /lei21/, the lowest tone, and /t^{h}iŋ55/, the highest tone in Cantonese tone system. By placing the target word in the middle of a sentence, intonation effects at the beginning and ending of an utterance can be minimized.

One clear utterance judged by the experimenter was selected for each talker. The target word /ji33/ produced by each talker was extracted and normalized in duration and intensity. Based on the measurement of the original production (mean = 460.96 ms, SD = 30.81), the duration of the target word produced by each talker was normalized to 500 ms in Praat (Boersma and Weenick, 2009). To balance the intensity level across four talkers, the peak intensity level of each target word was normalized to 60 dB. F0 and segmental cues of the target word were preserved.

To match the loudness level of the target word and its neighboring context, the average intensity level of all speech contexts was normalized to 60 dB, identical to the peak intensity level of the target word in PRAAT. The F0 of speech contexts was then manipulated to investigate how the relative F0 of the context affected tone perception. Following Wong and Diehl (2003), the overall F0 trajectory of each speech context was raised by two semitones, kept unshifted[2] and lowered by three semitones. The scale of the F0 shift was determined by the perceptual distance of three level tones in Cantonese. According to Chao (1947), high level tone is three semitones higher than mid level tone, which is in turn two semitones higher than low level tone. By shifting the contextual F0 in a contrastive way, an identical target word /ji33/ (mid level tone) is expected to be perceived as high level tone in the lowered F0 condition, as mid level tone in the unshifted F0 condition, and as low level tone in the raised F0 condition (Wong and Diehl, 2003). Manipulation of the contextual F0 shift gave rise to 12 sentences (4 talkers × 3 types of F0 shifts), which comprised the F0 contour condition of the speech context. As for the mean F0 condition, the original F0 contour of each speech context was replaced with the flattened F0 equal to the mean F0 of each context.

Afterward, the F0 trajectory and intensity profile of speech contexts were extracted from the F0 contour and flattened F0 conditions, respectively, and used to synthesize nonspeech contexts. This study used a triangle wave, which has a different harmonic structure from speech sounds, to generate nonspeech contexts. Also for the purpose of matching the loudness level of the target word and the context, the average intensity level of nonspeech contexts was set to 80 dB, 20 dB higher than the speech equivalents. It was rated by the first author and two naïve listeners that the nonspeech contexts (80 dB) sounded similar in loudness level as the speech contexts (60 dB).

After the manipulation, target words were embedded in the speech and nonspeech contexts in a talker-congruent way (i.e., /ji33/ produced by a talker was inserted in the contexts from the same talker). Figure 1 displays the spectrogram of one speech utterance and the nonspeech counterpart.

In addition to the test items, fillers produced by the same four talkers were included to avoid the expectancy effect of hearing the same sentence. One filler sentence, 我唔識意字點寫 /ŋo23 m21 sek5 ji33 tsi22 tim23 se25/ "I don't know how to write /ji33/" was recorded from two female talkers, and a second sentence, 我嚟讀意俾你聽 /ŋo23 lei21 tuk22 ji33 pei25 lei23 t^{h}iŋ55/ "I will read /ji33/ for you" was recorded from two male talkers. Following the procedures described in the preceding text, four types of contexts (speech and nonspeech contexts, and contexts with F0 contour and flattened F0) were generated for each filler sentence. But the F0 of the filler sentences was not shifted. The ratio of test items and fillers was 3:1.

C. Procedure

The stimuli were presented in five blocks. The first block contained the isolated target words, i.e., /ji33/ produced by all four talkers. This isolation condition ("isolated") served as the baseline for examining talker normalization without contextual cues. The other four blocks corresponded to the four context conditions, i.e., speech context with F0 contour ("sp_dy"), speech context with flattened mean F0 ("sp_mn"), nonspeech context with F0 contour ("ns_dy"), and nonspeech context with flattened mean F0 ("ns_mn"). Within one block,

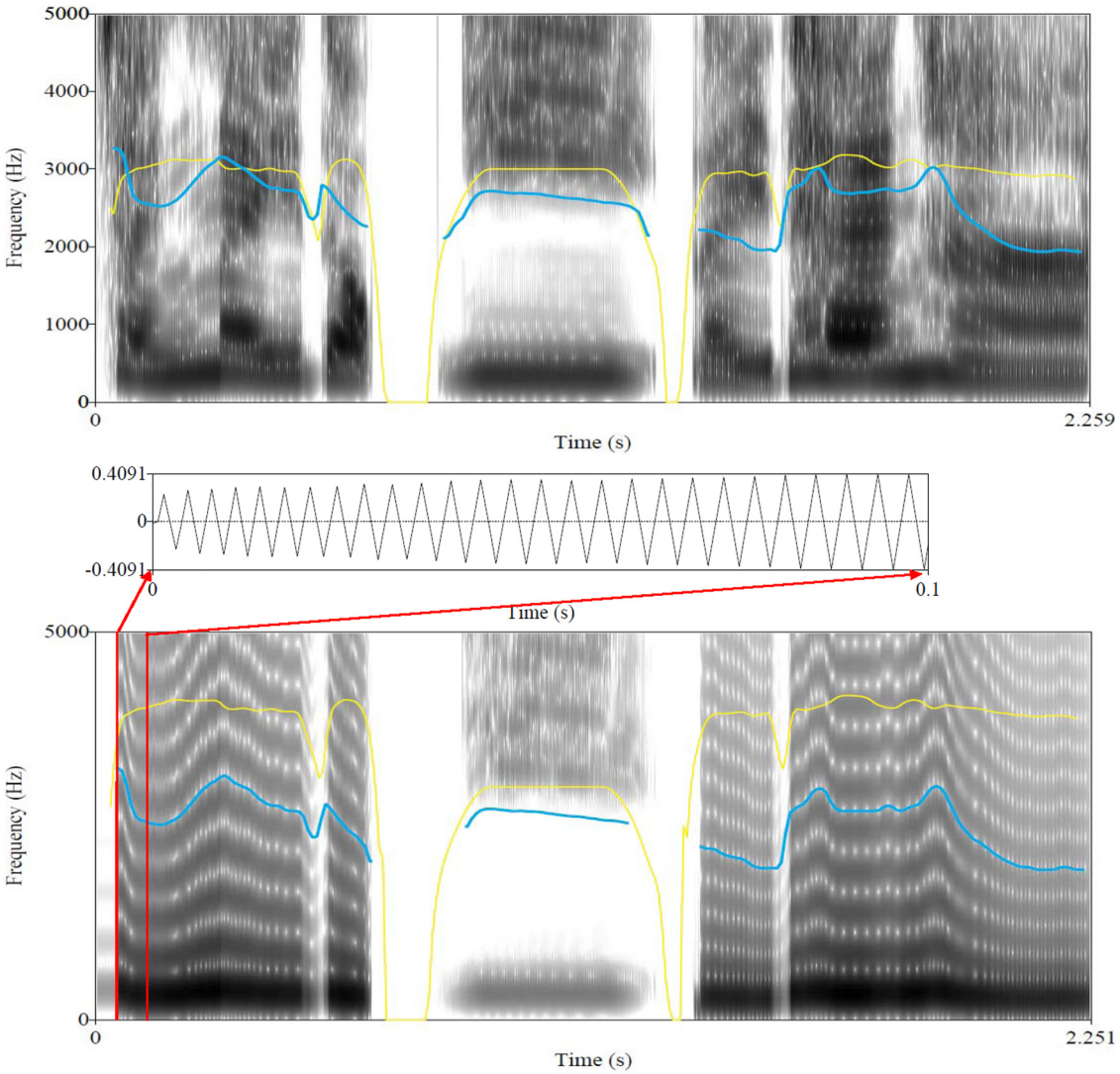

FIG. 1. (Color online) Spectrogram of a speech sentence and its nonspeech counterpart (synthesized with a triangle wave). The thick line represents the F0 contour, and the thin line displays the intensity profile. Waveform of a small portion of the nonspeech context is displayed above the spectrogram of the nonspeech context.

all 16 stimuli [(3 test items + 1 filler) × 4 talkers] was presented in random order and repeated for seven times. Across all five blocks, the target words were the same, i.e., /ji33/ produced by the four talkers, while the context condition varied from block to block.

The isolation block was always presented first. For the remaining four blocks, two nonspeech blocks always preceded two speech blocks. As discussed earlier, previous studies consistently confirmed the effect of speech context on tone normalization (e.g., Lin and Wang, 1984; Wong and Diehl, 2003; Francis *et al.*, 2006), whereas the effect of the nonspeech context was controversial (Francis *et al.*, 2006; Huang and Holt, 2009, 2011). To avoid the carryover of facilitation effect from speech contexts to nonspeech contexts, nonspeech blocks were presented before speech blocks. Within the speech and nonspeech blocks, the order of F0 contour and flattened F0 blocks was counterbalanced across the subjects.

Two practice blocks were presented to familiarize the subjects with the experiment procedure. The practice items were recorded from two talkers other than the preceding four talkers. One talker produced the test sentence and the other talker produced one of the filler sentences, which were processed following the procedure described in the preceding text. The first practice block, with isolated words, was presented before the isolation block. The second practice block, with target words embedded in the nonspeech context, was presented before the remaining four test blocks.

All the stimuli were presented at a sound pressure level that was comfortable for each subject. This level was determined for each subject during the first practice block and was kept constant within a subject. The task was three-alternative forced choice identification. For the isolation block, subjects were instructed to identify the target word as any of the three Cantonese words, 醫 (/ji55/ "a doctor"), 意 (/ji33/ "meaning"), and 二 (/ji22/ "two"). For the four context blocks, subjects were instructed to attend to the whole utterance and identify the target word after the whole utterance was presented. Subjects were asked to respond by pressing labeled buttons on a computer keyboard within three seconds.

D. Analysis

Two types of analyses were used in this study, perceptual height analysis and identification rate analysis. In the first analysis, each tone response was coded according to the predefined perceptual height value. Average perceptual height of all tone responses was calculated per contextual F0 shift (raised, unshifted and lowered) and per context condition (5 conditions) for each listener (Wong and Diehl, 2003; Francis *et al.*, 2006). As mentioned earlier, high level tone is three semitones higher than mid level tone, which is two semitones higher than low level tone (Chao, 1947). Following this perceptual scale, a high level tone response was coded as "6," a mid level tone response as "3," and a low level tone response as "1," with 6, 3, and 1 referring to a tone response's perceptual height (Wong and Diehl, 2003). This analysis had the advantage of estimating the overall change of tone responses according to the F0 shift in a context condition. If the average perceptual height was close to 1, it indicated that most stimuli were identified as low level tone in a condition. If it was close to 6, it indicated that most stimuli were identified as high level tone. It should be noted that when the value was close to 3, there was no unique inference about the ratio of tone responses. For example, the perceptual height of a condition where /ji33/ was mostly identified as mid level tone would be close to the perceptual height of a condition comprising equal portions of three tone responses. In such cases, the perceptual height analysis can be complemented by the identification rate analysis.

The second analysis calculated the identification rate of an expected tone response per contextual F0 shift (raised, unshifted and lowered) and per context condition (5 conditions) for each listener. Given the contrastive context effect reported before (Wong and Diehl, 2003; Francis *et al.*, 2006), the target word /ji33/ is expected to be perceived as high level tone in the lowered F0 condition, as mid level tone in the unshifted condition, and as low level tone in the raised F0 condition.

III. RESULTS

The first subsection reports the general picture of the effects of four context conditions as revealed by the perceptual height analysis. In the second subsection, the general picture is divided into three F0 shift conditions and each is examined with the identification rate analysis. The motivation for this decomposition is to zoom-in on the specific normalization patterns for different talkers in each F0 shift condition. As mentioned earlier, four talkers with different F0 ranges were recruited to introduce perceptual ambiguity into perception. If one type of context efficiently facilitates tone normalization, the identification rate should be higher than the chance level consistently for all four talkers. The third subsection provides a close examination of the effect of nonspeech contexts on the identification rate of tone stimuli.

A. General results

Figure 2 displays the percentage of three tone responses within each F0 shift condition (raised, unshifted, and lowered)

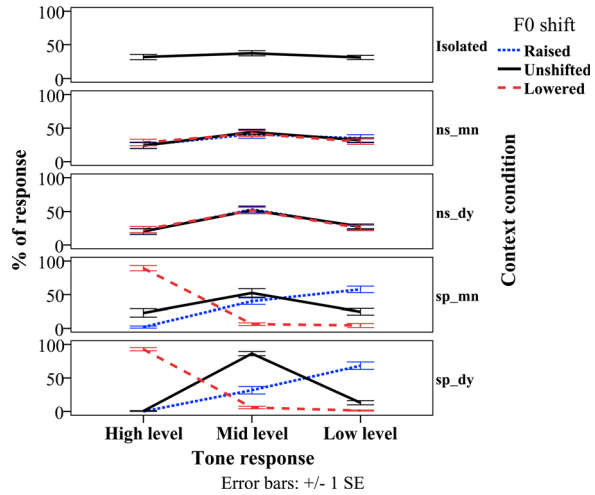

FIG. 2. (Color online) Ratio of three tone responses in the identification of /ji33/ according to the F0 shift in five context conditions (Isolated = no context, ns_mn = nonspeech context with mean F0, ns_dy = nonspeech context with F0 contour, sp_mn = speech context with mean F0, sp_dy = speech context with F0 contour).

in each context condition (5 conditions). Average perceptual height calculated for each condition is shown in Fig. 3.

A three-way repeated measures ANOVA was conducted on the perceptual height of four context conditions (the isolation condition was excluded due to the lack of F0 shift in this condition). *Speech type* (speech and nonspeech contexts), *F0 type* (contexts with F0 contour and flattened F0), and *F0 shift* (raised, unshifted, and lowered) were indicated as three within-subjects factors. Greenhouse–Geisser method was used to correct violations of sphericity where appropriate.

There were significant main effects of speech type, $F(1, 63) = 4.52$; $p < 0.05$ and F0 shift, $F(2, 126) = 696.42$; $p < 0.001$. Moreover, there were two significant two-way interactions—speech type by F0 shift, $F(2, 126) = 863.44$; $p < 0.001$, and F0 type by F0 shift, $F(2, 126) = 4.95$;

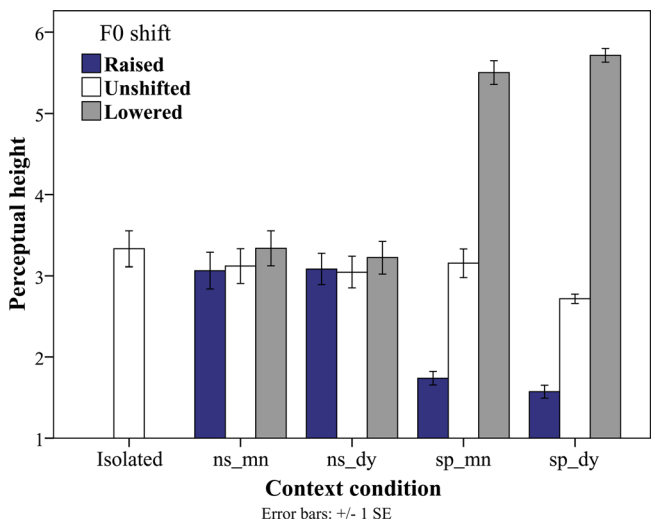

FIG. 3. (Color online) Perceptual height (see the text) of tone responses for three types of F0 shift in five context conditions. Isolated = no context, ns_mn = nonspeech context with mean F0, ns_dy = nonspeech context with F0 contour, sp_mn = speech context with mean F0, sp_dy = speech context with F0 contour.

$p < 0.01$. The three-way interaction also reached significance, $F(2, 126) = 7.06$; $p < 0.01$. It suggested that shifting the contextual F0 modified the perception responses to identical target words and that the influence of F0 shift was modulated by the speech type (speech and nonspeech) and F0 type (F0 contour and flattened F0) of the context conditions.

An indicator of the effect of a context condition on tone perception was how efficiently the relative F0 height of this context changed the perception responses. To further compare the effect of four context conditions, speech type × F0 shift and F0 type × F0 shift two-way repeated measures ANOVAs were then conducted (i.e., how F0 shift interacted with the speech type and F0 shift of the four contexts).

First of all, speech type × F0 shift repeated measures ANOVAs were conducted for the F0 contour and flattened F0 conditions separately. There were significant interaction of F0 shift by speech type in both the F0 contour contexts, $F(1.79, 112.91) = 825.10$; $p < 0.001$, and flattened F0 contexts, $F(1.73, 109.04) = 259.35$; $p < 0.01$. The results indicated unequal effects of speech and nonspeech contexts in general. Shifting the F0 in speech contexts (for both F0 contour and flattened F0 contexts) changed the perceptual height in a way that the target word /ji33/ (mid level tone) was mainly perceived as low level tone in the raised F0 condition (perceptual height close to 1) and mainly as high level tone in the lowered F0 condition (perceptual height close to 6). However, shifting the contextual F0 in nonspeech contexts showed no obvious effect on changing the tone perception, as the perceptual height value was close to 3 across three F0 shift conditions (see Fig. 3).

F0 type × F0 shift repeated measures ANOVAs were then conducted for the nonspeech and speech contexts separately. The interaction effect of F0 type and F0 shift only reached significance in the speech contexts, $F(2, 126) = 7.95$; $p < 0.01$. The results suggested that the original and flattened F0 conditions showed different effects on tone perception in the speech contexts but not in the nonspeech contexts. Comparing the F0 contour and flattened F0 conditions in speech contexts, Fig. 3 revealed that the perceptual height of the F0 contour condition was closer to 1 in raised F0 condition (1.57 vs 1.74) and that the perceptual height was closer to 6 in lowered F0 condition (5.71 vs 5.50). It implied that the F0 contour condition was more efficient in eliciting expected tone responses.

In summary, statistical analyses revealed unequal effects of speech and nonspeech contexts on tone perception, i.e., raising or lowering the F0 of speech contexts changed the tone perception in a contrastive way, whereas shifting the F0 of nonspeech contexts showed no obvious effect on the perception. Within the speech contexts the effects of the F0 contour and mean F0 conditions were also different. This difference is further explored in the following text.

B. Talker-based analysis of tone normalization

In this subsection, the results are divided into three F0 shift conditions and examined with the identification rate analysis. Results of the unshifted F0 condition are reported first, for the consideration that the expected tone response in the unshifted F0 condition was the original tone category of the target word (mid level tone). Examining the identification rate of the original tone category across four talkers allows us to first estimate the effect of different context on talker normalization. This subsection then proceeds to two shifted F0 conditions (raised and lowered) to report how efficiently F0 shifts in each context modified the perception of identical target words to high level or low level tone. Two-tailed *t*-tests were conducted to examine whether the identification rate of an expected tone was significantly higher than the chance level (33.33%) for each talker. Results of *t*-tests are summarized in Table I.

Figure 4 shows the identification rate of mid level tone (unshifted F0 condition) across four talkers in five context conditions. The isolation condition served as the baseline for examining the perceptual ambiguity induced by talker variability. Without contextual cues, listeners failed to correctly recognize the target word produced by all talkers except the female low (FL) talker. Further examination of the ratio of three tone responses within each talker (Table II) suggested that listeners were strongly influenced by a talker's F0 range. For example, listeners misidentified 42.9% of the /ji33/ tokens produced by the female high (FH) talker as *high level tone*, presumably because of the relative high F0 range of this talker. Although /ji33/ from the FL talker was mainly correctly recognized, 36.6% of the stimuli were still misidentified as *low level tone*. Perceptual confusion was more severe for two male talkers. 81.3% of /ji33/ produced by the male high (MH) talker was misidentified as high level tone, and 72.3% of /ji33/ from the male low (ML) talker was misidentified as low level tone. Different tendency of tone misidentification, high level tone for FH and MH and low level tone for FL and ML, suggested that tone perception was biased by the relative F0 range of two same-gender talkers.

The finding that Cantonese words produced by some talkers were easier to recognize (i.e., FL) is intriguing, which

TABLE I. *p* values of two-tailed *t*-tests that compared the identification rate of an expected tone response with the chance level (33.33%). N.S. = non-significant. *p* value in parentheses indicates that the identification rate is significantly *lower* than the chance level.

Lowered F0	FH	FL	MH	ML
ns_mn	N.S.	($p < 0.001$)	**$p < 0.05$**	($p < 0.001$)
ns_dy	N.S.	($p < 0.001$)	N.S.	($p < 0.001$)
sp_mn	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**
sp_dy	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**
Unshifted F0	**FH**	**FL**	**MH**	**ML**
Isolated	N.S.	**$p < 0.01$**	($p < 0.05$)	N.S.
ns_mn	**$p < 0.05$**	**$p < 0.01$**	N.S.	N.S.
ns_dy	**$p < 0.01$**	**$p < 0.01$**	N.S.	N.S.
sp_mn	**$p < 0.05$**	**$p < 0.001$**	N.S.	N.S.
sp_dy	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**
Raised F0	**FH**	**FL**	**MH**	**ML**
ns_mn	($p < 0.01$)	N.S.	($p < 0.01$)	**$p < 0.01$**
ns_dy	($p < 0.001$)	N.S.	($p < 0.001$)	**$p < 0.01$**
sp_mn	N.S.	**$p < 0.01$**	**$p < 0.01$**	**$p < 0.001$**
sp_dy	N.S.	**$p < 0.001$**	**$p < 0.001$**	**$p < 0.001$**

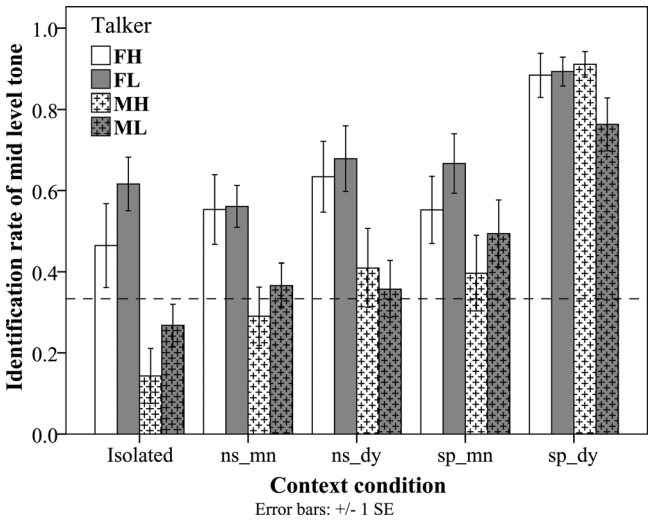

FIG. 4. Identification rate (see the text) of mid level tone in the unshifted F0 condition across four talkers. Isolated = no context, ns_mn = nonspeech context with mean F0, ns_dy = nonspeech context with F0 contour, sp_mn = speech context with mean F0, sp_dy = speech context with F0 contour. Dashed line indicates the chance level (33.33%).

may be related to Cantonese listeners' prior expectation of the average F0 range of talkers in the Cantonese-speaking community (Peng *et al.*, 2012). A talker whose F0 range was closer to the average (such as FL) may be easier to normalize even without contextual cues. This point is further discussed later.

TABLE II. Ratio of high level, mid level, and low level tone responses in the identification of /ji33/ stimuli produced by each of the four talkers in the unshifted F0 condition.

Isolated	High level tone (%)	Mid level tone (%)	Low level tone (%)
FH	**42.9**	**46.43**	10.7
FL	1.79	**61.61**	**36.6**
MH	**81.3**	**14.29**	4.46
ML	0.89	**26.79**	**72.3**
ns_mn	High level tone	Mid level tone	Low level tone
FH	**38.53**	**55.05**	6.42
FL	0.91	**56.36**	**42.73**
MH	**56.76**	**28.83**	14.41
ML	0	**36.61**	**63.39**
ns_dy	High level tone	Mid level tone	Low level tone
FH	**28.83**	**63.06**	8.11
FL	0	**67.86**	**32.14**
MH	**53.15**	**40.54**	6.31
ML	0	**35.71**	**64.29**
sp_mn	High level tone	Mid level tone	Low level tone
FH	**27.03**	**54.95**	18.02
FL	9.91	**66.67**	**23.42**
MH	**46.36**	**40.00**	13.64
ML	7.34	**48.62**	**44.04**
sp_dy	High level tone	Mid level tone	Low level tone
FH	**1.79**	**88.39**	9.82
FL	0	**89.29**	**10.71**
MH	**0.89**	**91.07**	8.04
ML	0	**76.15**	**23.85**

For the two nonspeech contexts and the speech context with flattened F0, the overall identification rate appeared to show some enhancement compared to the isolation condition (see Table II). However, identification rate for two male talkers failed to reach significance in any of these three context conditions. Even for the speech context with flattened F0, 46.36% of /ji33/ from MH talker was misidentified as high level tone, and 44.04% of /ji33/ from ML talker was misidentified as low level tone. It suggested that F0 cues of these three contexts were insufficient for normalizing the highly ambiguous word /ji33/ produced by two male talkers. Only the speech context with F0 contour consistently facilitated the normalization of all four talkers. It is interesting that speech context with flattened F0 was less efficient than the speech context with F0 contour in this condition.

Figure 5 illustrates the identification rate of high level tone (lowered F0 condition). In the two nonspeech contexts, although the identification rate appeared to be above chance level for the FH and MH talkers, only the identification rate for the MH talker in the nonspeech flattened F0 condition reached significance. In the two speech context conditions, the identification rate was significantly higher than the chance level for all four talkers.

Figure 6 displays the identification rate of low level tone (raised F0 condition). In the two nonspeech contexts, only the identification rate for ML was significantly higher than the chance level. In the two speech contexts, the identification rate reached significance in all talkers except for FH. Lack of significant effect for FH talker may be attributed to the influence of the high F0 range of this talker, which somehow resisted the effect of raising F0 in speech contexts.

In summary, results of the identification rate analysis confirmed unequal effects of speech and nonspeech contexts, i.e., only speech contexts consistently facilitated talker normalization across four talkers. Within the two speech contexts, the F0 contour and flattened F0 conditions showed similar effects in the raised and lowered F0 conditions;

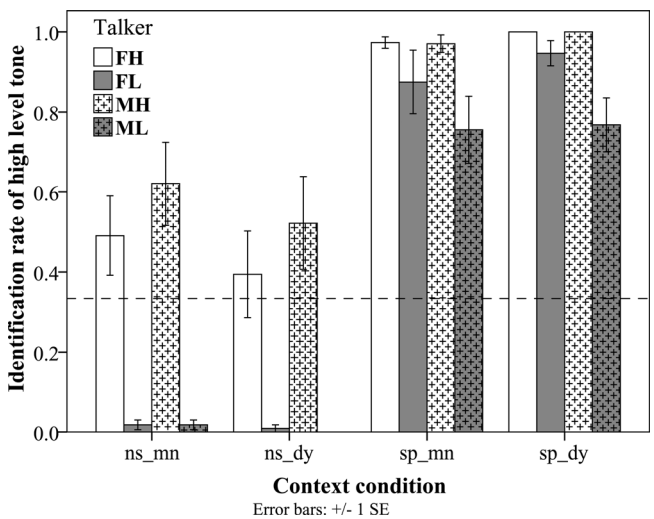

FIG. 5. Identification rate (see the text) of high level tone in the lowered F0 condition across four talkers. ns_mn = nonspeech context with mean F0, ns_dy = nonspeech context with F0 contour, sp_mn = speech context with mean F0, sp_dy = speech context with F0 contour. Dashed line indicates the chance level (33.33%).

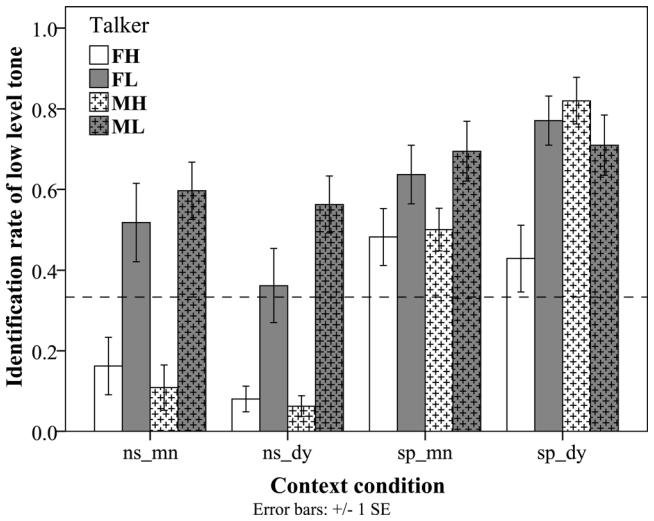

FIG. 6. Identification rate (see the text) of low level tone in the raised F0 condition across four talkers. ns_mn = nonspeech context with mean F0, ns_dy = nonspeech context with F0 contour, sp_mn = speech context with mean F0, sp_dy = speech context with F0 contour. Dashed line indicates the chance level (33.33%).

however, the flattened F0 condition failed to elicit reliable normalization in the unshifted F0 condition.

C. Close examination of the effect of nonspeech contexts

As mentioned earlier, Huang and Holt (2009, 2011) and Francis *et al*. (2006) reached different conclusions regarding the effect of nonspeech contexts partly because of the different analysis methods used. It is likely that the effect of nonspeech context failed to surface in Francis *et al*. (2006) due to the averaging of scores across all tone response. It is therefore worth carefully examining how the relative F0 height of nonspeech contexts modulated the ratio of each tone response. To this end, the identification rate of tone stimuli was calculated according to the congruent and incongruent F0 shift conditions. An F0 shift condition was defined as congruent if a tone response was expected given the contrastive contextual effect. For example, for high level tone response, the congruent condition was the lowered F0 condition. Accordingly, the incongruent conditions, which reflected the random response of high level tone irrespective of contextual cues, included the unshifted and raised F0 conditions. The purpose was to examine whether the congruent condition increased the ratio of a tone response over the incongruent conditions (averaged from two incongruent conditions). *Tone response* (high level, mid level, and low level) × *congruency* (congruent and incongruent) repeated measures ANOVA was carried out for the two nonspeech contexts, respectively.

For the flattened F0 condition, there was a significant main effect of congruency, $F(1, 15) = 7.72$, $p < 0.05$, suggesting that significantly more expected tone responses were elicited in the congruent F0 shift conditions (high level tone: 28.55% vs 24.17%; mid level tone: 44.22% vs 41.33%; low level tone: 34.67% vs 30.78%). For the F0 contour condition, the congruent condition also elicited more high level and low level tone responses than incongruent conditions (high level tone: 23.10% vs 20.15%; mid level tone: 51.84% vs 52.36%; low level tone: 26.96% vs 26.54%), but the difference was not statistically significant.

IV. DISCUSSION

A. The process of talker normalization in lexical tone perception

Tone normalization relies on both word-internal cues and contextual cues. Word-internal cues include F0, duration, intensity, voice quality, and other acoustic cues with F0 cues being the most important correlate for tone perception. Talker characteristics are also implemented in the word-internal F0 cues (Lee, 2009; Peng *et al*., 2012). This study found that the efficiency of normalization based on word-internal cues is talker-dependent. A word with mid level tone produced by some talkers is highly ambiguous without contextual cues (e.g., MH and ML), whereas the same word produced by other talkers can be correctly recognized (e.g., FL). This talker-dependent effect may reflect that listeners build prior expectations of average speaking F0 of talkers in a speech community (e.g. Honorof and Whalen, 2005; Lee, 2009; Peng *et al*., 2012). The words produced by a talker whose speaking F0 is close to the average of a speech community are likely to be correctly recognized without contextual cues. Accordingly, words produced by a talker whose speaking F0 is higher or lower than the average are likely to be biased toward words with high tones or low tones respectively. Peng *et al*. (2012) estimated the average F0 range of Cantonese male and female talkers from a speech database of 68 Cantonese talkers (34 female, 34 male) (Lee *et al*., 2002). Comparing the average F0 range of Cantonese male and female talkers (female: 200–290 Hz; male: 110–160 Hz) in Peng *et al*. (2012) with the four talkers recruited in the present study (female high: 180–350 Hz, female low: 180–280 Hz, male high: 110–190 Hz, male low: 80–130 Hz), it reveals that the F0 range of FL talker is largely overlapping with the female average F0 range, which then accounts for the result that the words produced by FL can be easily recognized without contextual cues. Moreover, although the lower bound of the F0 range of MH talker is aligned with that of the male average F0 range, the upper bound of MH talker is much higher. This discrepancy explains why most of /ji33/ stimuli from MH are misidentified as high level tone. Moreover, the direction of tone misidentification, high level tone for FH, and low level tone for ML, can also be accounted for along this line.

When the speaking F0 of a talker is unknown, prior knowledge about the average F0 range of Cantonese talkers may facilitate talker normalization (e.g., FL), because presumably there are more talkers whose speaking F0 are close to the average (Peng *et al*., 2012). But the prior knowledge also gives rise to the bias in tone perception, especially when a talker's speaking F0 is far away from the average (e.g., MH and ML).

In this sense, talker normalization is a process that listeners overcome the prior bias and tune to a particular talker's speaking F0 in speech perception. To minimize the prior

bias, the neighboring context that contains cues of a talker's speaking F0 is essential. It is likely that listeners build an expectation of a talker's speaking F0 from the context. By integrating the F0 cues from the context with the word-internal F0 cues, listeners are able to modify the prior bias, thereby correctly recognizing the words produced by different talkers. Moreover, speaking F0 varies within a talker (Garrett and Healey, 1987; Protopapas and Lieberman, 1997). Raising or lowering the overall F0 of a sentence produced by the same talker, reflecting the intra-talker variation, leads the listeners to update the expectation of this talker's speaking F0, which then elicits a change in the perception of identical target words. This process also implies that the mapping from the acoustic signals to phonological categories is not deterministic but probabilistic depending on all the available cues. In tone normalization, probabilistic mapping of acoustic signals (e.g., a level pitch) to competing phonological categories (e.g., high level tone, mid level tone, and low level tone) is evaluated against prior expectations that are built from the immediate context as well as from long term experience (i.e. expectation of average F0 range). A phonological category granted the highest probability in each condition is selected in the response. When the prior expectations are updated, the probabilistic mapping of acoustic signals to competing categories is also modified.

In summary, talker normalization involves the dynamic process of continuously updating the expectation of a talker's phonetic space from the context and interactively evaluating the incoming acoustic signals of a new word (i.e., the target word) against the expectation (e.g., Clark, 2012). This dynamic process proceeds along the incoming speech stream in speech perception.

B. Speech and nonspeech contexts

A fundamental question is asked at the beginning of this study: Is tone normalization mediated by a speech-specific process or a general perceptual process?

This study found that shifting the F0 trajectory in speech contexts changes the perception of identical target words contrastively, whereas nonspeech contexts show no obvious effect. It confirms the findings of Francis *et al*. (2006). Moreover, only speech contexts reliably facilitate the tone normalization for four talkers with different speaking F0. Nevertheless, it does not mean that nonspeech contexts play no role. Close examination of the effect of nonspeech contexts reveals that F0 shifts in nonspeech contexts also mildly shift the *identification preference* of target words in a contrastive way, similar to the effect reported by Huang and Holt (2009, 2011).

However, the effects of speech and nonspeech contexts are qualitatively dissimilar in this study. F0 shift in nonspeech contexts mildly shifts the identification preference of tone stimuli, but F0 shift in speech contexts changes the perceived tone category. It is likely that the findings of this study do not differ from Huang and Holt (2009, 2011) in terms of the effect of *nonspeech contexts*, but they do differ in terms of *speech contexts*. We suspect that the type of tone stimuli may have contributed to the different effects of speech contexts. Huang and Holt (2009, 2011) studied contour tones (specifically, a continuum of high level to high rising tone stimuli), whereas this study investigated level tones. Word-internal cues of contour tones are less ambiguous; this may have limited the effect of speech contexts (e.g., Peng *et al*., 2012). Word-internal cues of a level tone are ambiguous (i.e., a level pitch can be mapped to any of the three level tones in Cantonese); therefore shifting the F0 trajectory of speech contexts effectively changes the perception from one level tone to another level tone. This speculation warrants future experiments for verification.

Why do speech and nonspeech contexts show unequal effects? In speech communication, speech sounds that encode linguistic information are ecologically more important than nonspeech sounds for information transmission. It is likely that different neural networks and mechanisms have been evolved to process speech and nonspeech sounds (Whalen *et al*., 2006; Liberman *et al*., 1967; Liberman and Mattingly, 1985). The processing of speech sounds therefore is relatively immune to the influence of nonspeech sounds, which are processed by different networks.

It is also likely that attunement to the speaking F0 of a particular talker requires both segmental and suprasegmental cues. Although nonspeech contexts carry identical suprasegmental cues as speech contexts, there are no segmental cues. Without segmental cues, nonspeech contexts do not sound like the vocalization of any talker. It is likely that listeners consider nonspeech contexts to be irrelevant for estimating the speaking F0 of a talker. As a result, F0 cues of the nonspeech context do not affect the perception of the target word at large. The [ə] context used in Francis *et al*. (2006) falls short of facilitating the normalization presumably because the [ə] context does not sound like natural vocalization. To some degree, the difference between speech and nonspeech contexts may also be related the fact that our living environment is filled with various noises. In the so called "cocktail party" phenomenon (e.g., Bronkhorst, 2000), reliable talker normalization requires accurately tuning to a target talker's voice and filtering out irrelevant voices and nonspeech sounds. Nonspeech context may be ignored due to its irrelevance for tuning to a target voice.

In summary, tone normalization is likely to recruit the speech-specific mechanism (e.g., Francis *et al*., 2006; Liberman *et al*., 1967; Liberman and Mattingly, 1985), and general auditory mechanism only plays a marginal role. Future experiments are needed to examine whether the lack of semantic meanings in the nonspeech contexts have also contributed to the different effects of speech and nonspeech contexts.

C. F0 range and mean F0

The preceding discussion reveals that only speech contexts efficiently facilitate tone normalization. A question that follows is what kind of F0 cues in the speech context is extracted to estimate a talker's speaking F0. Listeners may rely on a talker's full F0 range or a talker's mean F0.

In the F0 contour condition, listeners can estimate both the full F0 range and mean F0 of a talker from the context.

But in the mean F0 condition where the F0 contour is neutralized, only the information of the mean F0 is available. The F0 contour condition reliably facilitates the normalization of both inter- and intra-talker variation (Wong and Diehl, 2003; Francis *et al*., 2006). Interestingly, the mean F0 condition works efficiently when the contextual F0 is raised and lowered; however, it fails to facilitate talker normalization in the unshifted F0 condition. The finding of similar effects of the original and mean F0 conditions in the *shifted F0* conditions is consistent with that of Francis *et al*. (2006). But in the *unshifted F0* condition, mean F0 cues alone are insufficient for resolving the high lexical ambiguity in the case of some talkers (i.e., MH and ML, see Table II), a finding not reported in the previous studies.

In the unshifted F0 condition, the mean F0 of the context is close to the mean F0 of the target word /ji33/, whereas in shifted F0 conditions (raised and lowered), the mean F0 of the context is far away from that of the target word. As discussed earlier, listeners may evaluate the word-internal F0 cues against the contextual cues. When the mean F0 of the context mismatches with (i.e., stays far away from) that of the target word, it may elicit the illusion that this particular talker has a higher-than-expected or lower-than-expected mean F0, which then leads the listeners to modify the perception of the target word in a way congruent with the relative F0 height of the context. However, when the mean F0 of the context matches with (i.e., stays close to) that of the target word, no such illusion is elicited. As a result, listeners may treat the F0 cues of the context as uninformative, which is then largely disregarded in the perception of the target word.

In the F0 contour condition, listeners may estimate both the F0 range and mean F0 cues of a talker from the context. Even if the mean F0 cue is not informative in the unshifted F0 condition, the F0 range of a talker allows the listeners to unambiguously compute the relative position of a level pitch within the F0 range, thereby accurately mapping it to the intended tone category.

In summary, mean F0 of the context (without explicit information of a talker's upper and lower F0 range) is less efficient in tone normalization. It is worth noting that the task difficulty is high in this study. These four talkers selected in this study have different F0 ranges, introducing great perceptual ambiguity into tone perception. The mean F0 of a talker is useful for tone normalization in most conditions. But a talker's full F0 range is more efficient for resolving lexical ambiguity, especially when the talker-induced lexical ambiguity is high. The effects of F0 range and mean F0 cues on talker normalization are worth further investigation.

ACKNOWLEDGMENTS

This work is supported in part by grants from the Research Grant Council of Hong Kong (GRF: Grant No. 455911), from National Natural Science Foundation of China (NSFC: Grant Nos. 11074267, 61135003), and a 973 Grant from the National Basic Research Program of the Ministry of Science and Technology of China (Grant No. 2012CB720700). We thank Dr. Hong-Ying Zheng, and Dr. James W. Minett for their valuable comments on this study.

[1]A talker's F0 range is estimated from his/her production of two Cantonese words in isolation, /ji55/ (high level tone) for the estimation of the upper bound and /ji21/ (low falling tone) for the estimation of the lower bound.

[2]For the unshifted condition, the F0 trajectory of a context was raised by 1 Hz. The purpose of raising the F0 in the unshifted conditions is to balance the possible artifact caused by F0 manipulation in the raised and lowered F0 conditions (Wong and Diehl, 2003). Such mild F0 shift in this condition is expected not to affect the perception of the target word.

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin), pp. 109–278.
Boersma, P., and Weenink, D. (**2009**). "Praat: Doing phonetics by computer (version 4.0) [Computer program]," http://www.praat.org (Last viewed February 21, 2012).
Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. **86**, 117–128.
Chao, Y.-R. (**1947**). *Cantonese Primer* (Harvard University Press, Cambridge, MA), pp. 1–242.
Clark, A. (**2012**). "Whatever next? Predictive brains, situated agents, and the future of cognitive science," Behav. Brain Sci. (Sec. 3), 1–86.
Fox, R. A., and Qi, Y.-Y. (**1990**). "Context effects in the perception of lexical tone," J. Chin. Linguist. **18**, 261–284.
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.
Holt, L. L. (**2006a**). "Speech categorization in context: Joint effects of nonspeech and speech precursors," J. Acoust. Soc. Am. **119**, 4016–4026.
Holt, L. L. (**2006b**). "The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization," J. Acoust. Soc. Am. **120**, 2801–2817.
Holt, L. L., and Lotto, A. J. (**2008**). "Speech perception within an auditory cognitive science framework," Curr. Dir. Psychol. Sci. **17**, 42–46.
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 (L)," J. Acoust. Soc. Am. **129**, 1145–1148.
Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 363–389.
Johnson, K., and Mullennix J. W. (**1997**). *Talker Variability in Speech Processing* (Academic, San Diego), pp. 1–237.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
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, C. Y., Tao, L., and Bond, Z. S. (**2009**). "Speaker variability and context in the identification of fragmented Mandarin tones by native and nonnative listeners," J. Phonetics **37**, 1–15.
Lee, T., Lo, W. K., Ching, P. C., and Meng, H. (**2002**). "Spoken language resources for Cantonese speech processing," Speech Commun. **36**, 327–342.
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.
Lin, T., and Wang, W. S-Y. (**1984**). "Shengdiao ganzhi wenti (Tone perception)," Zhongguo Yuyan Xuebao **2**, 59–69.
Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.
Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. M., and Wang, W. S-Y. (**2012**). "The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems," J. Speech Lang. Hear. Res. **55**, 579–595.

Protopapas, A., and Lieberman, P. (**1997**). "Fundamental frequency of phonation and perceived emotional stress," J. Acoust. Soc. Am. **101**, 2267–2277.
Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**, 81–110.
Rose, P. (**1996**). "Cantonese citation tones," in *Vocal Fold Physiology: Controlling Complexity and Chaos*, edited by P. J. Davis and N. H. Fletcher (Singular Pub. Group, San Diego), pp. 307–324.
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.
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. (**1998**). "Speaker normalization in the perception of Cantonese level tones," Master's thesis, University of Texas at Austin.
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.