Normalizing Talker Variation in the Perception of Cantonese Level Tones:

Impact of Speech and Nonspeech Contexts

Caicai Zhang, Gang Peng, William S-Y. Wang

Language Engineering Laboratory, the Chinese University of Hong Kong, Hong Kong, China

yzcelia@gmail.com, gpengjack@gmail.com, waywang@ee.cuhk.edu.hk

Abstract

Cantonese level tones (high level, mid level, and low level tone) are perceptually confusing if produced by talkers with different pitch ranges. External context with cues of a talker’s pitch range is essential for resolving the ambiguity of level tones. This study investigates the impact of two contexts – speech and nonspeech contexts on the perceptual normalization of talker variation in Cantonese level tones. Four talkers with different pitch ranges are asked to produce the stimuli so that perceptual ambiguity is expected without normalization. We find unequal effects of speech and nonspeech contexts. F0 cues in speech context are efficiently used as reference for normalizing talker difference in tone perception, whereas nonspeech context carrying identical F0 cues shows no obvious effect. It suggests that the carrier of F0 cues (speech or nonspeech) affects whether the context is engaged or not in tone normalization. This finding implies that tone normalization is likely to be a speech specific process, congruent with the phonetic module of speech perception.

Index Terms: talker normalization, tone perception, context, level tones, Cantonese

1. Introduction

In a speech community, people differ in the way they talk. The same word spoken by different talkers shows tremendous acoustic difference in the speech signals [1, 2]. Despite the talker variability in speech production, listeners can recognize a word spoken by different talkers without much difficulty. This phenomenon is known as talker normalization [1, 2]. Talker normalization serves as a basic mechanism for supporting efficient communication within a speech community. Humans’ ability to normalize talker variation is so powerful that we can understand what a stranger says from the very beginning of a conversation. Compared to the extraordinary ability of humans, computers often show lowered accuracy in automatic speech recognition if the speech signals are from multiple talkers (e.g. [3]). The ability of talker normalization in humans may have evolved for a long time, maybe together with the emergence of the spoken language. Such fast and efficient talker normalization in humans must have benefited from the long evolutionary history of the spoken language.

In a tone language, differences in fundamental frequency (F0) can be used to distinguish lexical meanings [4]. Due to the physiological differences in vocal apparatus, talkers differ in the speaking fundamental frequency, giving rise to variation in the F0 realization of the same tone (e.g. [5]). Speaking fundamental frequency also varies within the same talker across the day [6] and changes as a function of emotional mood [7] and other factors.

Despite the talker variation in tone production, listeners are capable of extracting invariant tone categories and therefore recognizing words. The process that listeners recognize the same tone produced by different talkers is referred to as ‘tone normalization’ (e.g. [8, 9]). Although this phenomenon is well recognized, its mechanism is not yet clear. According to the Pitch Range Assessment Model [9, 10], listeners may adapt to the pitch range of a talker, which serves as reference for judging the identity of a certain tone. Following this model, external context (i.e. context neighboring a target word) that contains cues of a talker’s pitch range is expected to be crucial for tone normalization [8, 9]. As far as speech context is concerned, previous studies consistently attested the effect of the external context on tone perception [8, 9, 11, 12, 13, 14, 15]. Although not all of these studies examined context effect in the multiple-talker condition, their findings bear general implication for talker normalization, which is elaborated below.

Lin and Wang [11] found that identical Mandarin target word attached to a word produced with either high or low pitch (i.e. different acoustic realizations of linguistically the same word) was perceived as different words. Although the target word was kept constant, its perception changed in a way contrastive to the relative F0 height of the context, i.e., the high-pitch context elicited more low tone responses whereas the low-pitch context elicited more high tone responses. It demonstrated the integration of the context and the target word in tone perception, i.e., the contextual F0 was used as reference for evaluating the tone identity of the target word. Although this study did not test talker variability, its finding is relevant for understanding talker normalization. As mentioned earlier, speaking fundamental frequency varies within a talker across the time. The immediate context may inform listeners of the dynamic change of a talker’s pitch range at one particular time. The context effect therefore showed that listeners rescaled the pitch percept of the target word according to the immediate context, which reflected the most updated status of F0 variation in a talker’s speech.

Leather [12] and Moore and Jongman [13] explicitly examined the effect of speech context on the perceptual normalization of talker variability. They found that embedding talker ambiguous Mandarin tone stimuli in speech utterances produced by two different talkers led to the perception of different tones, depending on which of the two talkers was perceived to ‘produce’ them. In other words, listeners adjusted the pitch percept of Mandarin tone stimuli according to the immediate context.

Similar results had been reported in Cantonese. Francis et al. [8] found that shifting the F0 trajectory of a Cantonese sentence while keeping the target word constant changed the perception of the target word in a contrastive way. Wong and Diehl [9] introduced talker variability into investigation. A Cantonese sentence “/ha22 jie5 koa33 tsii22 hoa22 si33/” [The...
next word is try’ (the last word /si33/ was the target word) was recorded from seven male Cantonese talkers. Wong and Diehl [9] found strong context effect, i.e., rising and lowering the F0 trajectory of the context changed the identification of the target word from mid level tone to low level tone (95.8%) and to high level tone (99.5%) respectively.

Consistent findings in different tone languages suggest that tone normalization in general involves the recruitment of F0 cues in speech context for normalizing talker difference. That F0 is the major acoustic correlate of tone raises a question as to whether identical F0 cues carried by speech and nonspeech contexts have equal effect on tone perception. With regard to nonspeech context, inconsistent findings were reported. Huang and Holt [14, 15] found that nonspeech context (harmonic tones and pure tone contexts) modeled after the average F0 of the speech context had a qualitatively similar, though reduced effect as speech context on the perception of Mandarin tones. However, Francis et al. [8] reported. Huang and Holt [14, 15] found that nonspeech context (harmonic tones and pure tone contexts) modeled after the average F0 of the speech context had a qualitatively similar, though reduced effect as speech context on the perception of Mandarin tones. However, Francis et al. [8] found that [ə] sound synthesized with the dynamic F0 trajectory of speech context had no effect on the perception of Cantonese level tones, unlike its speech equivalent. Despite some uncertainties regarding the interpretation of the [ə] sound – whether it is nonspeech sound or nonnative speech sound, [8] casts some doubt on the potential effect of nonspeech context on tone normalization.

Inconsistent results in the literature motivate the present study to further explore the effect of speech and nonspeech contexts. Moreover, none of the previous studies [8, 14, 15] have explicitly examined the effect of nonspeech context in the multiple-talker condition. To complement previous studies, the present study aims to thoroughly examine the effect of speech and nonspeech contexts in the multiple-talker condition. Cantonese level tones provide an ideal window to study this question. There are three level tones in Cantonese, high level, mid level, and low level tone (c.g. [19]). An isolated level pitch produced by different talkers is ambiguous in that it may be mapped to any of these three tones [9, 20]. Cantonese level tones therefore allow us to examine whether identical F0 trajectory superimposed on speech and nonspeech contexts have similar effect on tone normalization.

The present study follows the experiment design of [8, 9] at large, but complements the previous study [8] by synthesizing the nonspeech context with triangle wave that can be unambiguously perceived as nonspeech sounds. Results of the present study can confirm whether speech and nonspeech contexts have equal effect on tone normalization. This comparison may shed light on the nature of tone normalization (i.e. context effect) – whether tone normalization is a speech specific process [8, 16, 17] or general auditory process [14, 15, 18]. If only speech context facilitates tone normalization, it suggests that only speech sounds are useful for normalizing talker difference in tone perception, which is congruent with the view that only speech sounds are processed via the phonetic module [16, 17]. But if both speech and nonspeech contexts equally facilitate tone normalization, it supports a general perceptual mechanism which underlies the processing of both speech and nonspeech sounds [18].

2. Methodology

2.1. Participants

16 native speakers of Hong Kong Cantonese (8F, 8M; age=22.6±2.6) without hearing impairment or long-term music training were paid to participate in the experiment. All participants gave informed consent in compliance with a protocol approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong.

2.2. Stimuli

Four native Cantonese talkers (2F, 2M) were recruited to record the speech materials. Two talkers within each gender had different but overlapping pitch ranges so that misidentification of level tone was expected without talker normalization. Pitch range of these four talkers was estimated from their production of the highest and the lowest tone in Cantonese tone inventory – high level tone (医 /i55/ ‘a doctor’) and low falling tone (亞 /i21/ ‘a son’) in isolation [20].

Table 1. Approximate pitch range of four native Cantonese talkers.

<table>
<thead>
<tr>
<th>Talker</th>
<th>Lower range</th>
<th>Upper range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female High (FH)</td>
<td>180 Hz</td>
<td>350 Hz</td>
</tr>
<tr>
<td>Female Low (FL)</td>
<td>180 Hz</td>
<td>280 Hz</td>
</tr>
<tr>
<td>Male High (MH)</td>
<td>110 Hz</td>
<td>190 Hz</td>
</tr>
<tr>
<td>Male Low (ML)</td>
<td>80 Hz</td>
<td>130 Hz</td>
</tr>
</tbody>
</table>

To prepare the perceptual stimuli in the experiment, each talker was asked to read aloud a sentence 請你讀意願下/sin25 lei23 tuk22 i33 ln21 t/i55 ha23/ ‘please read /i33/ for me’ and repeat it for six times. 意/i33/ (mid level tone) ‘meaning’ in the middle of this sentence was the target word, and the remaining of this sentence served as the context. The context included words with high level tone (i.e. 嘅/ /i55/ ‘to listen’) and low falling tone (i.e. 嘅/i21/ ‘to come’), which provided cues of the full pitch range of a talker.

One clear sentence judged by the experimenter was selected for each talker for the following manipulation. The target word, /i33/, was normalized in duration and intensity profile across four talkers. The duration of the target word was kept constant at 500 ms, and the maximal intensity at 60 dB. Talker-specific F0 and segmental cues were preserved.

In order to investigate how F0 cues in the neighboring context affects the perception of the target word, the overall F0 trajectory of each sentence, excluding the target, was raised by two semitones, kept unshifted and lowered by three semitones [9]. The scale of F0 shift was determined by the perceptual distance of Cantonese level tones – high level tone is three semitones higher than mid level tone, which is in turn two semitones higher than low level tone [21]. By shifting the contextual F0 in a contrastive way, identical target word /i33/ (mid level tone) was expected to be perceived as high level tone in the lowered F0 condition, as mid level tone for the unshifted condition, and as low level tone for the raised condition respectively.

In addition to test sentences, filler sentences were included to avoid the expectancy effect of hearing the same sentence in the experiment. One filler sentence, 我唔識意願學寫/po23 m21 sek5 i33 ts22i tim23 se25/ ‘I don’t know how to write /i33/’ was recorded from two female talkers, and the other sentence, 我嚟個你讀 /po23 ln21 tuk22 i33 pei25 lei23 t/i55/ ‘I will read /i33/ for you’ was recorded from two male talkers. The ratio of test sentences and filler sentences was 3:1.

Nonspeech context was synthesized with triangle wave, which has rich harmonics but no segmental cues. F0 trajectory
and intensity profile that were extracted from speech contexts (both test and filler sentences) were used to synthesize the triangle wave contexts. The overall intensity level was set to be 20 dB higher in the triangle wave contexts than the speech contexts in order to balance the loudness level of speech and nonspeech contexts. Speech and nonspeech contexts were matched in duration and F0 trajectory.

**2.3. Procedure**

There were three blocks in total. The first block included isolated target words extracted from each sentence. This isolation condition served as the baseline for investigating talker normalization without context [9, 20]. The other two blocks were speech context condition and nonspeech context condition respectively. Across these three blocks, the target words were the same, i.e., /i33/ produced by four talkers, but the context conditions were different among blocks, i.e., no context (isolation block), speech context (speech block) and nonspeech context (nonspeech block). Within each block, all the stimuli were randomized and repeated for seven times.

In each trial, listeners were instructed to identify the target word as any of these three Cantonese words, 医 (i55/, high level tone, ‘doctor’), 医 (i33/, mid level tone, ‘meaning’) and 僺 (i22/, low level tone, ‘second’) by pressing labeled buttons on a computer keyboard. Listeners were given 3 seconds to respond after the whole utterance was presented.

These three blocks were presented in this order: (1) isolation block, (2) nonspeech block, and (3) speech block. The presentation order was determined based on previous findings of the effect of speech and nonspeech contexts. Nonspeech context has been found to affect tone perception inconsistently [8, 14, 15]. As a result, nonspeech block was presented after the isolation block in order to avoid any potential transfer effect from nonspeech block to isolation block. Moreover, previous studies have consistently found that listeners adapt to a talker’s pitch range through familiarization with the speech context [9, 12, 13], the speech block was presented at last.

**2.4. Analysis**

Two types of analyses were used in this study, perceptual height analysis and identification rate analysis. The first analysis was to calculate the average perceptual height of all identification responses in a condition. Following Wong and Diehl [9], each identification response was coded according to the perceptual height of the selected tone. As mentioned earlier, high level tone (T1) is three semitones higher than mid level tone (T3), which is in turn two semitones higher than low level tone (T6) [21]. Following this perceptual scale, each high level tone response was coded as ‘6’, each mid level tone response as ‘3’, and each low level tone response as ‘1’, with ‘6’, ‘3’ and ‘1’ referring to the perceptual height of a tone (see Figure 2). After coding each tone response according to its perceptual height, the average perceptual height of all responses was calculated for three F0 shift conditions (raised, unshifted and lowered) within each context condition (no context, speech context and nonspeech context) for each listener. If the average perceptual height of responses in a F0 shift condition was close to ‘1’, it indicated that most target words in this condition were perceived as low level tone. If it was close to ‘6’, it indicated that most target words were perceived as high level tone.

The second analysis was to calculate the identification rate of the expected tone response in a condition. According to the contrastive context effect reported before (i.e. more low tone responses if the contextual F0 is high and vice versa) [8, 9, 11, 12, 13], the target word /i33/ was expected to be perceived as high level tone for the lowered F0 context, as mid level tone for the unshifted condition, and as low level tone for the raised condition. The identification rate of expected tone responses was calculated for three F0 shift conditions (raised, unshifted and lowered) within each context condition (no context, speech context and nonspeech context) for each listener.

The advantage of the perceptual height analysis is that it estimates the change of the overall tone responses according to the F0 shift. However, when the perceptual height of a condition was close to ‘3’, there was no unique inference about the composition of tone responses (e.g. the perceptual height of a condition comprising one third of each tone response would be roughly ‘3’, close to a condition in which identification responses were mostly mid level tone). Nevertheless, this analysis can be complemented by the identification rate analysis when the average perceptual height was close to ‘3’. Both analyses were used in this study to provide a fuller picture of the context effect on normalization.

**3. Results**

**3.1. Perceptual height analysis**

Figure 3 displays the average perceptual height of responses in three F0 shift conditions within nonspeech and speech contexts. A two-way repeated measures ANOVA was conducted by indicating Context type (speech and nonspeech contexts) and F0 shift (raised, unshifted, and lowered) as two within-subjects factors. Greenhouse-Geisser method was used to correct violations of sphericity where appropriate.

There was significant main effect of F0 shift (F(2, 126) = 618.006, p < 0.001), but no significant main effect of Context type (F(1, 63) = 2.464, p = 0.121), suggesting an overall difference in perceptual height among three types of F0 shift, but not between speech and nonspeech conditions. Moreover, there was significant interaction of Context type by F0 shift (F(2, 112.908) = 825.103, p < 0.001). This significant interaction was intriguing, which indicated unequal effects of speech and nonspeech contexts on tone normalization. Post-hoc one-way ANOVAs comparing three F0 shift conditions (raised, unshifted, and lowered) in speech and nonspeech contexts separately revealed that the unequal effects were due...
to significant difference among three F0 shifts in speech context \((F(2, 165) = 817.860, p < 0.001)\), but not in nonspeech context \((F(2, 165) = 0.227, p = 0.797)\). It suggested that only shifting the F0 trajectory of the speech context led to the change of tone perception. As shown in Figure 3, perceptual height of the raised F0 condition in the speech context was close to ‘1’, suggesting that the target word was mostly identified as low level tone \((T6)\) in this condition; perceptual height of the lowered F0 condition was close to ‘6’, suggesting that the target word /i33/ was predominantly identified as high level tone \((T1)\). However, raising or lowering the F0 trajectory of the nonspeech context showed no obvious effect of changing the tone perception.

Figure 3: Average perceptual height of tone responses in three contextual F0 shifts of speech context and nonspeech context conditions.

In summary, listeners rescaled the pitch percept of the target word according to the F0 cues of the speech context, but F0 cues of the nonspeech context was not efficiently used to evaluate the tone identity of the target word.

3.2. Identification rate analysis

This subsection reports the identification rate of expected tone response for all four talkers (Female High, Female Low, Male High, and Male Low) in speech and nonspeech contexts. As mentioned earlier, two talkers within each gender had different but overlapping pitch ranges so that misidentification was expected without normalization. If contextual cues efficiently facilitated the normalization, it meant that the identification rate of expected tone responses should be higher than the chance level not only in one or two talkers, but consistently in all four talkers. Examining the identification rate across all four talkers can further reveal the strength of the speech and nonspeech contexts in talker normalization. To this end, two-tailed t-tests were conducted to compare the identification rate of expected tone responses with the chance level accuracy (33.3%) for each talker in speech and nonspeech contexts. T-test results were summarized in Table 2.

Table 2. Summary of two-tailed t-test results. Asterisk indicates that the identification rate was significantly higher than the chance level. ‘*’ = 0.05 < p < 0.01; ‘**’ = 0.01 < p < 0.001; ‘***’ = p < 0.001. ‘N.S.’ indicates lack of significant effect.

<table>
<thead>
<tr>
<th>Tone Level</th>
<th>Female High (FH)</th>
<th>Female Low (FL)</th>
<th>Male High (MH)</th>
<th>Male Low (ML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High level</td>
<td>N.S.</td>
<td>**</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Speech</td>
<td>**</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Mid level</td>
<td>N.S.</td>
<td>**</td>
<td>N.S.</td>
<td>N.S.</td>
</tr>
<tr>
<td>Speech</td>
<td>**</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Low level</td>
<td>N.S.</td>
<td>N.S.</td>
<td>N.S.</td>
<td>**</td>
</tr>
<tr>
<td>Speech</td>
<td>N.S.</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

Figure 4 shows the identification rate of high level tone in the unshifted F0 condition. Since the expected tone response in the isolated target word condition (isolation block) was mid level tone, this condition was included here. The isolated condition served as the baseline for talker normalization without context. Figure 5 shows that only in the speech context the mid level tone was correctly recognized consistently across all four talkers. In the isolated condition, only the target word produced by FL talker was correctly identified. In the nonspeech context, the overall identification rate showed mild enhancement compared to the isolated
condition, but still the identification rate failed to reach significance for two male talkers.

Figure 6 represents the identification rate of low level tone in the raised F0 condition. In the speech context condition, the identification rate was significantly higher than the chance level in all talkers except FH. Lack of significance for FH talker may be attributed to the influence of the high pitch range of this talker, which is discussed later in this subsection. In the nonspeech context condition, although the identification rate was significantly above chance level in the ML talker, it failed to reach significance in the remaining three talkers.

![Figure 6: Identification rate of low level tone (T6). Dashed line represents the chance level (33.3%).](Image 63x496 to 277x644)

Across the board, only in the speech context condition, the identification rate of expected tone response was above the chance level consistently for all four talkers. Although nonspeech context carried identical F0 cues as the speech context, the identification rate failed to reach significance for two or more talkers in the nonspeech condition. If efficient talker normalization implies that a word produced by any talker should be correctly identified, only the speech context reliably facilitated the normalization of talker variation.

Without adapting to the pitch range of a particular talker through external context, listeners were likely to be biased by the pitch range of different talkers, which was revealed by the patterns of misidentification in the nonspeech condition. Here we limited our discussion to the unshifted F0 condition, as the misidentification pattern was similar among the raised, unshifted and lowered F0 conditions in nonspeech context (which again suggested limited effect of nonspeech context on tone normalization).

The above discussion revealed some interesting talker-specific normalization patterns, implying that listeners may resort to the expected pitch range of Cantonese talkers for normalization [20]. Speech context facilitates talker normalization in most conditions. But if a talker’s pitch range is much higher than the expectation, mid level tone produced by this talker tended to be biased towards high level tone despite the facilitation of the speech context (e.g. FH). Meanwhile, if a talker’s pitch range is largely overlapping with the expected range, the target word can be correctly recognized even without contextual cues (e.g. FL). Talker-specific normalization patterns suggest that contextual cues and the expected pitch range co-contribute to tone normalization in an interactive way.

### Table 3. Ratio of high level tone (T1), mid level tone (T3), and low level tone (T6) responses in the unshifted F0 condition of nonspeech context.

<table>
<thead>
<tr>
<th>Talker</th>
<th>T1</th>
<th>T3</th>
<th>T6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female High (FH)</td>
<td>28.8%</td>
<td>63.1%</td>
<td>8.1%</td>
</tr>
<tr>
<td>Female Low (FL)</td>
<td>0%</td>
<td>67.9%</td>
<td>22.1%</td>
</tr>
<tr>
<td>Male High (MH)</td>
<td>33.2%</td>
<td>40.5%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Male Low (ML)</td>
<td>0%</td>
<td>35.7%</td>
<td>64.3%</td>
</tr>
</tbody>
</table>

As shown in Table 3, for two female talkers, the target word /i33/ were mainly correctly identified as mid level tone, but a small portion of trials were misidentified, which reflected the influence of relative pitch height of these two talkers. For example, 28.8% of /i33/ word produced by FH talker was misidentified as high level tone, but only 8.1% were misidentified as low level tone. On the contrary, 32.1% of /i33/ word produced by FL talker was misidentified as low level tone, and none as high level tone. Misidentification was more serious in the case of two male talkers. 53.2% of /i33/ from the MH talker was misidentified as high level tone. Different from MH talker, most of the misidentification for the ML talker was due to low level tone (64.3%).

Different direction of tone misidentification, high level tone for FH and MH, and low level tone for FL and ML, suggested that tone perception was affected by the relative pitch of two talkers within each gender.

Indeed, the influence of pitch range differences between talkers was so strong that it mildly affected tone normalization even in the speech context condition. As mentioned earlier, in the raised F0 condition, the identification rate failed to reach significance for the FH talker. The relative high pitch range of FH talker to some degree prevented the target word from being perceived as low level tone (the expected tone response) in this condition. In addition, the identification rate of ML talker appeared to be lower than the other three talkers in the lowered and unshifted F0 conditions (but still significantly higher than the chance level), which was also due to the influence of the relative low pitch range of this talker.

The above discussion revealed some interesting talker-specific normalization patterns, implying that listeners may resort to the expected pitch range of Cantonese talkers for normalization [20]. Speech context facilitates talker normalization in most conditions. But if a talker’s pitch range is much higher than the expectation, mid level tone produced by this talker tended to be biased towards high level tone despite the facilitation of the speech context (e.g. FH). Meanwhile, if a talker’s pitch range is largely overlapping with the expected range, the target word can be correctly recognized even without contextual cues (e.g. FL). Talker-specific normalization patterns suggest that contextual cues and the expected pitch range co-contribute to tone normalization in an interactive way.

### 4. Discussion

Tone perception is not merely based on the pitch of a word alone, but is sensitive to acoustic cues in the neighboring context. As mentioned earlier, speaking fundamental frequency differs between talkers and also within a talker across the time, giving rise to overlapping in the F0 realization of different tones (e.g. [5]). To resolve the lexical ambiguity caused by talker variation, neighboring context is indispensable due to the fact that context contains cues of the dynamic organization of the pitch range of one particular talker at one particular time. It has been argued that speech perception may not involve many-to-one mapping from various acoustic realizations to one unique linguistic unit, but one-to-one mapping from each acoustic realization to talker-and context-specific representation of that unit (e.g., [22, 23]). Systematic phonetic details of talker- and context-specific representation are crucial for recovering the encoded information in speech perception [22,23,24]. Following this view, ‘talk normalization’ may not refer to a process of reduction from talker variation to formal linguistic units, but instead, dynamic adaptation to rich talker- and context-specific variation of linguistic sounds through the continuous integration and updating of contextual cues.

This context-dependent way of perception may not be unique to the auditory domain. There is some parallel between talker normalization in speech perception and the phenomenon of size constancy in visual perception. An object
is perceived as the same object despite the change of its appearance as a function of its relative position to the viewer. Perceptual constancy in both auditory and visual domains may involve the dynamic computation of a certain object relative to the neighboring context.

Returning to tone perception, an interesting question is what kind of context is useful for perceiving the relative pitch of a word in tone perception. That F0 is the most important acoustic correlate of tone perception gives rise to this question: do identical F0 cues embodied in different contexts (speech or nonspeech) affect tone perception in a similar way? In this study, we found that only the speech context reliably facilitated the normalization of talker variation, which is consistent with Francis et al. [8]. Nevertheless, it does not mean that nonspeech context plays no role in tone normalization [14, 15]. As shown in Figure 5, nonspeech context mildly increased the overall identification rate of mid level tone compared to the isolated condition. It suggests that nonspeech context may still be useful, but its effect is marginal compared to that of the speech context.

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 cortical networks or mechanisms are evolved to process speech and nonspeech sounds [16, 17, 25], which may have contributed to the difference here. Moreover, to adapt to the dynamic pitch range of a particular talker, listeners may require both segmental and suprasegmental cues to identify a talker. Without segmental cues, as in the case of the nonspeech context, there is insufficient information to ensure that the target word and the context are from a single talker. As a result, F0 cues of the nonspeech context do not affect the perception of the target word at large. To some degree, this difference may also be related the fact that our living environment is filled with various noises. In so called ‘cocktail party’ phenomenon (e.g. [26]), 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.

Findings of the present study lead us to draw the interim conclusion that tone normalization is more likely to be a speech specific process than a general auditory process [8], congruent with the phonetic module of speech perception [16, 17, 25]. A pending question is what acoustic cues in the speech context are indispensable for triggering the speech-mode of perception. Moreover, future studies may examine the time course of the context-target integration via electrophysiological studies, which will shed more light on the cortical mechanism of tone normalization.

5. Acknowledgements

This work is supported in part by grant 455911 from the Research Grants Council of the Hong Kong SAR. We thank Dr. Tao Gong, Dr. James Minett, and Dr. Hong-Ying Zheng for their constructive comments in improving the paper, and all other members at the Language Engineering Laboratory of the Chinese University of Hong Kong.

6. References