Extrinsic context affects perceptual normalization of lexical tone

Alexander L. Francis[a)]
Department of Speech, Language and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana 47907

Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu
Division of Speech and Hearing Sciences, Faculty of Education, 5th Floor, Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong

(Received 28 April 2005; revised 7 October 2005; accepted 11 October 2005)

The present study explores the use of extrinsic context in perceptual normalization for the purpose of identifying lexical tones in Cantonese. In each of four experiments, listeners were presented with a target word embedded in a semantically neutral sentential context. The target word was produced with a mid level tone and it was never modified throughout the study, but on any given trial the fundamental frequency of part or all of the context sentence was raised or lowered to varying degrees. The effect of perceptual normalization of tone was quantified as the proportion of non-mid level responses given in F0-shifted contexts. Results showed that listeners' tonal judgments (i) were proportional to the degree of frequency shift, (ii) were not affected by non-pitch-related differences in talker, (iii) and were affected by the frequency of both the preceding and following context, although (iv) following context affected tonal decisions more strongly than did preceding context. These findings suggest that perceptual normalization of lexical tone may involve a "moving window" or "running average" type of mechanism, that selectively weights more recent pitch information over older information, but does not depend on the perception of a single voice. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2149768]

PACS number(s): 43.71.Bp, 43.71.−k, 43.71.An, 43.71.Es, 43.71.Hw [ARB] Pages: 1712–1726

I. INTRODUCTION

Listeners may perceive the same acoustic pattern as different phonemes depending on the phonetic context it appears in (Liberman *et al.*, 1967), the speaking rate at which it appears to be produced (Verbrugge *et al.*, 1976), and the talker that is perceived to have produced it (Ladefoged and Broadbent, 1957). Such context-dependent processing of speech can be referred to as (perceptual) normalization, or more specifically according to the nature of the contextual information being used (e.g., phonetic normalization, speaking rate normalization, talker normalization). Note that this use of the term normalization may be distinguished from another use referring to the physical transformation of a signal to reduce between-token variability (e.g., peak amplitude normalization, or acoustic normalization more generally). The two uses of the term are historically related, in the sense that early conceptualizations of perceptual normalization assumed that listeners were mentally transforming incoming signals to derive context-independent representations in the same way that an engineer might use acoustic normalization to transform a physical signal to eliminate context-specific variability (see Johnson, 1997 for discussion). The assumption that perceptual normalization must result in a context-independent mental representation of the signal is not, however, a necessary one for the purposes of the present discussion (Johnson, 1997; Nusbaum and Magnuson, 1997). Here we will refer to talker normalization or tone normalization in keeping with the established literature on the topic, though we use these terms to refer simply to the process by which listeners understand speech (and lexical tones in particular) in a talker-dependent manner.

Talker normalization has been demonstrated for vowels (Ladefoged and Broadbent, 1957; Nearey, 1989), consonants (Johnson, 1991), and lexical tone (Jongman and Moore, 2000; Moore and Jongman, 1997; Wong, 1998; Wong and Diehl, 2003). For example, Ladefoged and Broadbent (1957) showed that a synthetic vowel in a "b_t" context was perceived as "bit" in isolation but as "bet" when preceded by a precursor sentence (Please say what this word is:) that was synthesized to have a generally lower first formant (F1) frequency. In other words, listeners interpreted the "bit" token as having a comparatively higher F1 when it was presented in the context of a sentence with overall lower F1 values.

The above-noted examples show that perceptual normalization, particularly talker normalization, derives at least in part from information provided by "extrinsic context" (or "extrinsic information")—speech that does not constitute part of the syllable or phoneme to be identified. Such extrinsic information seems to be particularly significant in the case of perception of lexical tones. In a lexical tone language, meaning can be distinguished according to the suprasegmental feature of tone alone. In Cantonese, the primary physical correlate of tone is fundamental frequency (f0) (Fok Chan, 1974; Vance, 1976). Thus, two syllables with different f0 contours may have different meanings, even when their

a)Electronic mail: francisa@purdue.edu

 0001-4966/2006/119(3)/1712/15/$22.50

segmental content is the same. For example, in Cantonese the segmental string /ji/ means "doctor" when produced with a high level tone but it means "two" when produced with a low level tone.[1] What is particularly interesting about the Cantonese tonal system is that it contrasts three level tones (high, mid, and low, or 55, 33, and 22) that differ minimally in terms of their contours (Bauer and Benedict, 1997; Rose, 2000). As we shall see, this system means that talker normalization on the basis of extrinsic context plays a particularly significant role in the accurate perception of Cantonese tones.

Since most research on tone normalization has been conducted using Mandarin listeners and stimuli, it is instructive to first review these studies in order to better understand the phenomenon in question. Standard (Beijing) Mandarin has four tone categories: a high (55) tone (tone 1), a rising (25) tone (tone 2), a dipping (214) tone (tone 3), and a falling (51) tone (tone 4). Although Mandarin tones differ primarily in terms of f0 contour, they can also be affected by their perceived relative height. For example, Leather (1983) showed that the perception of syllables with f0 contours lying toward the middle of a continuum between that of a 25 (tone 2) and 55 (tone 1) tone were interpreted differently depending on the context in which they appeared.

Tone normalization has been studied in Mandarin by Leather (1983), Lin and Wang (1985), Fox and Qi (1990), and Moore and Jongman (1997). Leather (1983) embedded tonally ambiguous tokens in carrier sentences produced by talkers with very different average f0 ranges, and showed that listeners categorized stimuli with identical f0 according to the speaker characteristics provided by the preceding context. Both Lin and Wang (1985) and Fox and Qi (1990) examined the effect of manipulating the f0 of one syllable in two-syllable sequences. Lin and Wang (1985) examined the effect of manipulating the f0 of the second syllable on perception of the tone of the first syllable, while Fox and Qi (1990) examined the effect of manipulating the f0 of the first syllable on perception of the tone of a following syllable. Although the results of Lin and Wang (1985) provided stronger evidence for contextual normalization of tone, Fox and Qi (1990) also found some evidence to support the idea that tone identification is affected by extrinsic context. Taken together, the results of all of these studies strongly suggest that, when the f0 of the target syllable is held unchanged, the perception of its tone can be influenced by the f0 of a neighboring syllable. Moore and Jongman (1997) provided the clearest demonstration of this phenomenon to date. They showed that a given (synthesized) Mandarin Chinese syllable could be identified as having either of two different tones depending on the fundamental frequency (f0) of the preceding sentence.[2] When the context f0 was low, the target was identified as having a mid rising tone (tone 2). When the context f0 was higher, the identical target syllable was identified as having a low falling-rising tone (tone 3). Thus, even though Mandarin tones may be identifiable on the basis of other (non-f0) properties (Fu *et al.*, 1998; Whalen and Xu, 1992), there is a clear and well-established effect of the fundamental frequency properties of preceding speech on the perception of Mandarin lexical tones.

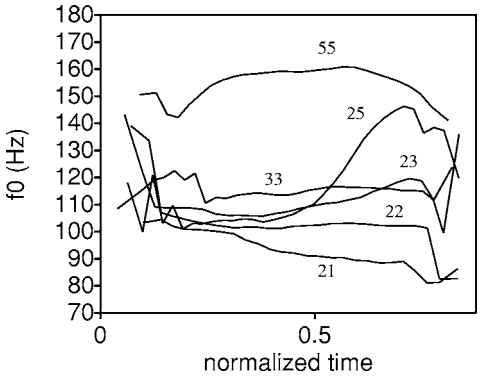

FIG. 1. Fundamental frequency (f0) contours for the six Cantonese tones produced by the speaker on whose speech the stimuli for experiment 1 were modeled.

Unlike Mandarin, Cantonese has three level tones that differ only in terms of their relative (average) f0 (Bauer and Benedict, 1997; Matthews and Yip, 1994; Rose, 2000; Vance, 1976; Wong, 1998; Wong and Diehl, 2003) as well as three contour tones (see Fig. 1). Thus, it is entirely possible that a low level tone produced by a talker with a very high average f0 might easily fall within the range of a high level tone produced by a talker with an overall lower average f0. This makes it imperative that listeners be able to judge tone within the context of the talker's individual f0 characteristics. Research by Wong (1998) and Wong and Diehl (2003) demonstrated the importance of extrinsic context in this process. They showed that, in Cantonese, it is possible to embed a syllable with a mid level tone (33) in a variety of contexts differing only in average f0, and have that tone be interpreted as either high, mid, or low level depending on the perceived f0 of the context. Wong (1998) and Wong and Diehl (2003) made use of this effect and manipulated the average fundamental frequency of a sentence preceding a target syllable produced with a mid level (33) tone. They showed that shifting the frequency of the preceding sentence upward by two semitones caused Cantonese listeners to hear the target as having a low level tone, while shifting the precursor frequency downward by three semitones caused the target to be identified as having a high level tone. Wong (1998) showed that this effect could be obtained not only with a Cantonese precursor, but also with an English precursor produced by the same talker, although the English effect was comparatively smaller. Since the English phonological system does not make use of lexical tones, these results suggested that Cantonese listeners did not need to access information specific to linguistically defined tonal categories in the precursor sentence. Rather, they could perform the task on the basis of nonlinguistic pitch properties alone. Wong (1998) observed that the overall pitch range of the talkers was smaller in English than in Cantonese. He argued that this suggested that Cantonese listeners were basing their tonal judgments on the position of the target syllable's f0 relative to the maximum (or minimum) pitch of the preceding context (the Pitch Range Assessment Model).

A. The Pitch Range Assessment Model (PRAM)

According to the Pitch Range Assessment Model (PRAM), listeners estimate a talker's tonal range on the basis

of the actual pitch ranges they experience in a given utterance. Therefore, a precursor sentence with a wider pitch range (the Cantonese precursor) should provide a more accurate estimation of expected tone productions than does one with a more constrained range (the English precursor). Similar arguments have been made regarding perceptual normalization of vowel spaces (Joos, 1948, p. 61). While these have generally been superseded by more sophisticated models, the reason for this seems to be at least in part because talker-specific vowel identification is quite accurate even in isolated syllables (Verbrugge *et al.*, 1976; see Johnson, 2005 for discussion). As we shall see here, although Cantonese tone identification is possible at better-than-chance levels in isolated syllables (Francis *et al.*, 2003; Wong, 1998; Wong and Diehl, 2003) and even without segmental information (Fok Chan, 1974), there is a substantial effect of extrinsic context, especially compared to that observed by Verbrugge *et al.* (1976) for vowels (see also Francis *et al.* 2003). Thus, a model such as PRAM may be more plausible in the case of tone normalization than it might be for vowel normalization. On the other hand, the PRAM was proposed on the basis of a very small number of experiments, and its predictive power remains limited. Many factors governing the operation of the PRAM remain to be determined.

B. The importance of pitch range

PRAM suggests that listeners will perform best when the context provides the widest possible frequency range for a given talker. Wong's (1998) results support this hypothesis, but could also be explained by listeners' different expectations regarding the role of pitch in English versus Cantonese rather than by any effect of pitch range per se. If bilingual Cantonese and English listeners expect f0 to play a different role in English than in Cantonese, they may rely less on contextual information about f0 when judging Cantonese syllables produced in an English context as compared to when they appear in a Cantonese context.[3] Thus, the apparent effect of talker pitch range may be spurious. Instead of using a relatively complex and highly variable measure such as pitch range over the course of a short, recent utterance to estimate a talker's tonal space, listeners might accomplish tone normalization on the basis of some more basic property of the talker's voice, for example average pitch, or even some non-pitch property of speech, perhaps one relating to overall vocal tract size (as has also been suggested for vowel normalization, see Johnson (2005) for details). To further explore this possibility, in experiment 1 we examined the response patterns of listeners hearing only one language (to eliminate effects of listeners' language-based expectations), and instead artificially varied the pitch range provided in the context.

C. Degree of shift

A second potential problem for the PRAM derives from the way stimuli were constructed. Wong (1998) and Wong and Diehl (2003) showed that a two-semitone upward shift, or a three semitone downward shift, were sufficient to induce tone normalization. These shift values were chosen because Chao (1947) reported that the Cantonese low level tone was approximately two semitones lower than the mid level tone, which was in turn approximately three semitones lower than the high level tone. However, data presented by Rose (2000; personal communication) suggests that the actual difference in modern Hong Kong Cantonese is approximately 4.1 semitones between the high level and mid level tones, and 1.6 semitones between the mid level and low level tones. Thus, it appears that Wong (1998) and Wong and Diehl (2003) may not have provided their listeners with an optimal stimulus configuration, yet their listeners still showed strong evidence for perceptual normalization of tone when stimuli were presented in context. This suggests that listeners do not need the pitch of the target syllable to lie exactly where it would be expected to be on the basis of the rest of the talkers' tone space, but instead accept a relatively wide range of possible frequencies as representative of each lexical tone. To explore this question, in experiment 1 (and subsequently) we also investigate the effect of the degree of contextual shift on tone normalization.

D. Preceding or following context

Third, with the exception of a single study by Lin and Wang (1985), most previous studies have used target syllables in utterance-final position. Thus, in most cases, contextual information was only provided prior to presentation of the target syllable. While such stimuli clearly provide listeners with optimal conditions for performing tone normalization, they cannot provide crucial information regarding the processing mechanisms involved. Most of the results presented thus far are consistent with a purely "feed-forward" model of tone normalization in which (only) previously occurring pitch information is used to estimate the tone category of a given syllable. However, hypothesizing such a model immediately raises the question of how listeners might judge the tone category of the first (or only) syllable in an utterance—a task that is certainly possible, but not well-studied. The results of Lin and Wang (1985) suggest that following context can be sufficient to facilitate tone normalization in the absence of a preceding context, but, because they used only a following context, their results still do not provide much insight into the relative weighting of preceding versus following context in tone normalization. In experiment 3 we explore the relative weighting of preceding and following context in tone normalization.

E. Talker specificity

Finally, previous studies have all implicitly assumed that tone normalization is a phenomenon depending primarily on fundamental frequency, regardless of listeners' perception of who is talking or what language they are speaking. Wong (1998) showed that listeners can accept changes in language, albeit with a slight reduction in the strength of normalization, suggesting that tone normalization is not completely language-specific. Fox and Qi (1990) have also proposed that tone normalization may derive from the operation of a mechanism that is not specific to linguistic processing. They showed that English and Mandarin speakers were equally

influenced by the F0 of a preceding syllable when asked to judge how similar a second syllable was to either Mandarin tone 1 (high level) or tone 2 (rising). Since English listeners were able to perform the task without any linguistic knowledge of the tonal system of Mandarin, Fox and Qi (1990) concluded that listeners depended solely on the acoustic information, i.e., the F0 of the context, to execute tone normalization. Thus, normalization was argued to be an auditory process, rather than a phonetic one.

In further support of the hypothesis that tone normalization does not require listeners to perceive the speech as being produced by a single talker, both Moore and Jongman (1997) and Leather (1983) have shown that tone normalization continues to operate despite (presumed) changes in voice quality within the same sentence. Both studies used synthetic target stimuli embedded within natural context sentences and showed normalization effects, which would suggest that some divergence in voice quality between context sentence and target syllable is acceptable to listeners. However, this aspect of tone normalization has not yet been investigated explicitly. If tone normalization results from the operation of a general auditory process rather than as an aspect of a language-specific process of talker normalization, then we would expect it to be robust even in the face of clearly noticeable changes in linguistic information and talker identity. We explore these factors in the second half of experiment 3 and in experiment 4.

In summary, the PRAM represents a preliminary model of tone normalization. However, tone normalization is not yet well understood, and many aspects remain to be investigated. Here we present the results of four experiments that, taken together, provide more insight into the operation of tone normalization, and suggest ways in which the PRAM can be augmented in order to account for these new phenomena.

II. EXPERIMENT 1

According to the PRAM, lexical tone normalization should be more effective when the extrinsic context provides the fullest possible range of frequency variation for a given talker. That is, if tone normalization depends on the range of frequencies in the context, then normalization should not take place if the context is completely monotone. Alternatively, tone normalization may involve the derivation of expected frequency values or ranges for particular tones based on a more abstract property of the extrinsic context, for example average f0. In this case, tone normalization should function equally well with a sentence with naturally varying frequency as with a sentence with a monotone f0 pattern that has the same average f0 as the natural one. To test this question, two different types of extrinsic context were used, one with a normal range of frequency variation, and one (monotone) in which the f0 of the context phrase was held constant at a value equal to the mean f0 over the other sentence (excluding the target).

Moreover, the original formulation of the PRAM implies that listeners' context-derived expectations of tone category locations should reflect experience with the typical (relative) f0 values of the tones as produced in the ambient language. Thus, for speakers of a language where the high level tone is typically produced at a frequency three semitones higher than the mid level tone, hearing a particular talker would induce a mental representation of that talker's tone space in which the high level tone is expected to be three semitones higher than the mid level tone. However, the degree of precision of such expectations is unknown. Is it sufficient that a token merely be higher in frequency than the expected frequency of a mid level tone in order for that token to be heard as a high level tone, or must the f0 value of the target syllable match the expected f0 of the high level tone category more precisely? Would a syllable with an f0 only two semitones higher than the expected mid level tone (or one semitone lower than the expected high level tone) still be heard as having a high level tone? In order to explore the relationship between the degree of frequency shift and the operation of tone normalization, three different degrees of pitch shift were used (one half semitone, one semitone, and two semitones).

A. Method

1. Participants

Twenty native Cantonese speakers (ten women and ten men) with no reported speech or hearing disability were recruited for the experiment. The mean age of the participants was 21.7 (range=20–24). Ten (five women and five men) were students in Speech and Hearing Sciences at the University of Hong Kong, with some training in phonetics. The remaining participants (five women and five men), were recruited from other faculties (schools) of the University of Hong Kong (Science, Arts, Law, Business, and Dentistry) and had no training in phonetics.

2. Stimuli

A Cantonese sentence: /ŋɔ23 wui23 tɔk22 ji33 pɛi25 lɛi23 t^hɛŋ55/我會讀意俾你聽 "I will read ji3 for you" modeled after the natural production of a native Cantonese male speaker (aged 22) was synthesized using Sensyn, a Klatt-style formant synthesizer from Sensimetrics Corporation (Klatt, 1980; Klatt and Klatt, 1990). To generate the synthetic stimulus, the natural sentence was sampled at 10 ms intervals and measurements of f0, amplitude envelope, and the first four formant frequencies were recorded and used as input to the Klatt synthesizer. A sentence was chosen that had the target syllable in the middle of the carrier phrase to avoid any potential for interaction with intonational effects on f0 related to the beginning or end of a phrase (cf. Vance, 1976). Six additional versions of the sentence were synthesized, with the nontarget portions of the sentence raised or lowered by $\frac{1}{2}$, 1, or 2 semitones while the target syllable itself remained unchanged (retained its original, natural f0 value of an average of 115 Hz). This resulted in a total of seven different carrier phrases, each surrounding a single syllable. These were designated as stimuli in the dynamic context condition. In addition, a monotone context version was created for each of the seven dynamic sentences by setting f0 to a uniform level equal to the average of the whole of each individual sentence (e.g., 116.0 Hz for the

TABLE I. F0 of each context sentence in the monotone context and average f0 of each context sentence in the dynamic context (experiment 1).

Direction of shift	Degree of shift	F0 (Hz)
Raised	2 semitones	131.1
	1 semitone	122.8
	$\frac{1}{2}$ emitone	119.4
Unshifted	0 semitones	116.0
Lowered	$\frac{1}{2}$ semitone	112.6
	1 semitone	109.2
	2 semitones	100.9

unshifted sentence). Thus, there were fourteen sentences in all: seven dynamic sentences (three with f0 raised by $\frac{1}{2}$, 1, or 2 semitones and three with lowered f0 (again, by $\frac{1}{2}$, 1, or 2 semitones) and one with an unshifted dynamic context), and seven monotone sentences corresponding to each of the dynamic sentences. Following Wong (1998) and Wong and Diehl (2003), in the raised conditions the target was expected to be identified as having a low level tone, while in the lowered conditions the target was expected to be identified as having a high level tone. Table I shows the f0 of each stimulus in the monotone context, calculated from the average f0 of the corresponding dynamic sentences.

3. Procedures

The experiment was carried out in a single-walled IAC sound-attenuating booth and took approximately 20 min to run. Stimuli were presented to listeners through Sennheiser HD-545 headphones, connected to an Apple PowerMacintosh 7100 computer. A Hypercard program was used for running the experiment. Before hearing the stimuli, an experimenter introduced and read aloud the three possible responses /ji55/醫 (doctor), /ji33/意 (*meaning*) and /ji22/ =(two) to the listeners, to ensure they were familiar with them.[4] A single block of fourteen stimuli was presented eleven times to each listener, for a total of 154 trials. Each block of trials contained the seven sentences from both the monotone and dynamic stimulus sets. All stimuli were presented in random order within each block.

In each trial listeners first heard a single sentence. Then the three possible responses were displayed on the screen above three numbered buttons. Listeners were asked to identify which of the three words appeared in the stimulus sentence by clicking one of the three buttons. The first block of trials (the first 14 trials) was treated as practice and results were not analyzed, although listeners did not know this at the time of the experiment.

B. Results and discussion

Shifting the fundamental frequency of a context sentence changed listeners' identification of a target syllable in the predicted manner: Downward shifts resulted in more high level responses, while upward shifts resulted in more low level responses. In order to more easily make comparisons across shift conditions, each response was scored according to whether it was the expected response for that condition (lowered, normal, and raised). For example, a response of /ji55/ (high level) was expected in the lowered context, while /ji22/ and /ji33/ were the expected responses in the raised and unshifted context conditions, respectively. The mean number of expected responses for each condition for each particular listener were then calculated and used for further analysis. No difference was found between participants with phonetics training and those without (mean =72% expected responses for both) or between male (mean=71%) and female participants (mean=72%). Therefore, these groups were combined for all further analyses.

Across all conditions, the degree to which responses matched predictions based on direction of shift was roughly proportional to the amount of shift (large, 2 semitone shift =100% expected response; medium, 1 semitone shift=74%; small, $\frac{1}{2}$ semitone shift=32%).[5] Two semitone shifts resulted in perfect or near perfect performance in the predicted direction, suggesting that listeners' expectations for tone locations in pitch space may be somewhat broader than might be predicted on the basis of Wong (1998) and Wong and Diehl's (2003) assumption, based on Chao's (1947) work, that a three semitone downward shift was necessary to induce listeners to hear the target as a high level tone. That is, while listeners may expect a high level tone to be approximately three semitones higher than a midlevel tone, they were equally willing to accept a tone that is only two semitones higher as a good high level percept. However, this is a tentative conclusion because of the presence of a ceiling effect, and the absence of some estimate of relative goodness.

Because of the ceiling effect in the two semitone pitch shift condition, the two semitone condition was excluded from further analysis. The number of expected responses for the unshifted condition was also quite high (not unexpectedly, mean=91% in monotone, and 94% in dynamic conditions). Therefore, this condition was also excluded from the analysis to simplify the data analysis. The proportion of expected responses for 1 and $\frac{1}{2}$ semitone shifts in both directions of both monotone and dynamic contexts is shown in Fig. 2.

A three-way within-subjects analysis of variance with factors context (monotone versus dynamic), direction (lowered versus raised), and size (one semitone versus one half semitone) was carried out. As expected, results showed a significant effect of size of shift, $F(1,19)=236.04$; $p<0.001$, such that there were more expected responses in the one semitone pitch shift condition (74%) compared to the half semitone shift condition (32%). There was also a significant effect of context, $F(1,19)=27.85$; $p<0.001$, such that the number of expected responses in the monotone context condition (mean=76%) was greater than in the dynamic context condition (mean=68%). This strongly suggests that listeners do not require prior experience with the complete range of a particular talker's pitch, but rather are able to compute the expected locations in pitch space of each lexical tone (at least the level ones) on the basis of the talker's average f0 alone. Similarly, there was a significant effect of direction, $F(1,19)=23.51$; $p<0.01$, such that there were fewer expected responses in the lowered condition (mean =59%) than the raised condition (77%). This was expected

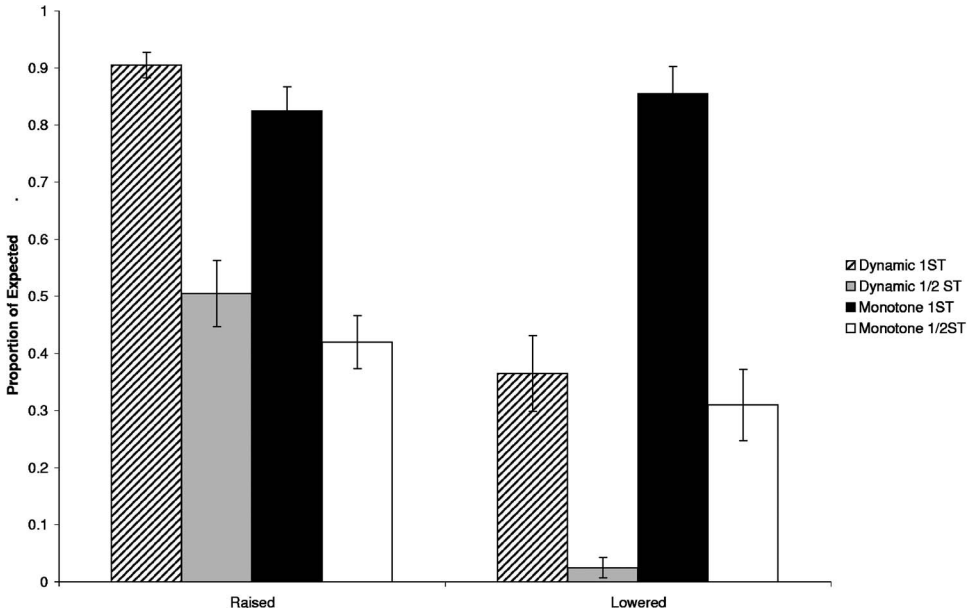

FIG. 2. Mean percentage of expected responses of 1 semitone and $\frac{1}{2}$ semitone shifts for monotone and dynamic contexts. Error bars indicate standard error.

because the two-semitone distance between a low level and a midlevel tone means that the maximum effect of raising would be reached within two semitones of upward shifting. In contrast, the *three-semitone* distance between midlevel and high level tones means that the maximum effect of lowering would only be reached with *three* semitones of (downward) shifting. Therefore, all else being equal, the raised condition should show more expected responses than the lowered condition, and this was indeed the case overall, especially for the dynamic context stimuli (71% expected response in the raised condition versus 20% in the lowered).

Unlike the dynamic condition, performance in the monotone condition was not affected by the direction of f0 shift (62% and 58% expected responses, respectively, and this was not significant by Tukey HSD post-hoc analysis, $p > 0.05$). The differential effect of monotone and dynamic f0 contours on the direction of F0 shift was supported by a significant interaction between context and direction, $F(1,19)=61.18$; $p<0.001$. Post hoc (Tukey HSD) analysis of this interaction revealed that all pairwise contrasts showed a significant difference ($p<0.05$) except for the one between the raised and lowered conditions in the monotone context, and the one between the dynamic and monotone contexts in the raised condition. This pattern of results suggests that something about the monotone context actually *improves* the overall effect of lowering the f0 of the context sentence. It is possible that the "robot-like" timbre of the synthetic speech might have encouraged listeners to treat the monotone (also "robot-like") context as somehow more acceptable, and therefore more effective, than the more dynamic context.

III. EXPERIMENT 2

The Pitch Range Assessment Model implies that optimal tone normalization can only be accomplished when the target syllable is preceded by a context containing speech at the upper and/or lower ends of the talker's pitch range. Experiment 1 of the present study demonstrated that it is not necessary to experience the full range of a talker's f0 range; exposure to the average f0 may suffice. Moreover, previous research suggests that tone normalization can exploit pitch information from either preceding (Wong, 1998; Wong and Diehl, 2003) or following (Lin and Wang, 1985) contexts, or both (the present experiment 1). However, it is still not clear how the temporal relationship between the context and the target affects tone normalization. Experiment 2 was designed to examine the relative importance of preceding versus following sentential context in lexical tone normalization.

A. Method

1. Subjects

Twelve college-aged, native Cantonese speakers (3 men and 9 women) with normal hearing and no history of speech or language disorder participated in this study.

2. Stimuli

The same semantically neutral sentence used in experiment 1 [/ŋɔ23 wui23 tɔk22 ji33 pɛi25 lɛi23 t^hɛŋ55/ "I will read *ji* for you (to hear)"] was recorded by a male native Cantonese volunteer in an IAC single walled sound-attenuating booth. Recordings were made via a Macintosh external microphone and recorded directly to disk using Sound Scope 16 (GW Instruments) via the built-in sound card of an Apple Macintosh G3 at a sampling rate of 44.1 kHz. Using Praat 3.9.27 (Boersma and Weenink, 2001), the stimulus was first low-pass filtered at 8k Hz. Subsequently, the f0 of the preceding and following contexts (those parts of the context sentence either preceding or following the word [ji]) was either (i) raised by one semitone, (ii) lowered by one semitone, or (iii) kept unshifted. Based on the results of experiment 1, a one-semitone shift was selected in order to avoid ceiling or floor effects in identification. All possible combinations of raising, lowering, and nonshifting were created, resulting in a total of nine conditions (3 levels of preceding context $\times 3$ levels of following context).

3. Procedure

The experiment was conducted in an IAC single-walled sound booth. The order of the nine stimuli in each block was randomized and 11 blocks were presented in all, resulting in a total of 99 trials for each participant. A Hypercard stack was used to present stimuli and collect responses. On each trial, a single stimulus was played out at 44.1 kHz via Sennheiser HD-545 headphones connected to an Apple PowerMacintosh 7100 computer, and the participant was asked to identify the target word by clicking on one of the characters shown on a computer screen. The possible characters were the same as in experiment 1. The first block served as a practice block and was excluded from analysis although the participants were not informed of this at the time. Each experiment session lasted for approximately 20 min.

B. Results

Participants' responses were recorded and scored with a "count-sum" scoring system: If the response indicated that the participant heard a low level tone then the response was given a score of −1; mid level tone responses were scored as 0; and high level tone responses were given a score of +1. The scores for each stimulus for each subject were summed. This resulted in a bounded range of scores from −10 to +10 since there were 10 trials for each stimulus. The more negative the score, the more the low level tone dominated the responses. Similarly, the more positive the score, the more the high level tone dominated the responses. If the score was close to 0, either the mid level tone dominated the responses or the three kinds of responses appeared roughly equally, or there were just high level and low level tone responses in almost equal measure. This method of scoring was chosen instead of the proportion of predicted responses used in experiment 1 because of the difficulty of determining *a priori* what the predicted response would be in cases where the preceding and following contexts were shifted in opposite directions. Figure 3 shows the scores for all the conditions.

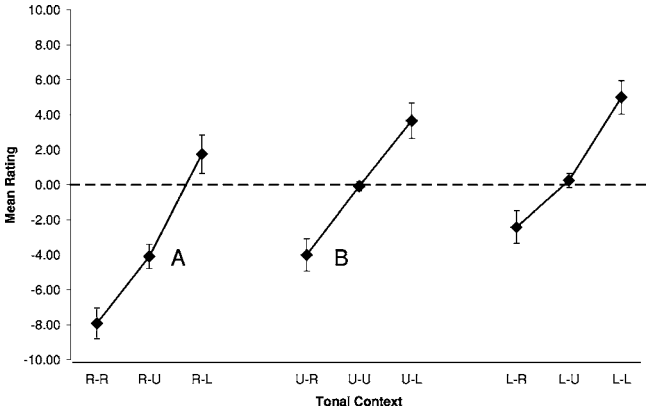

FIG. 3. Mean response scores (see the text) for tokens presented in contexts with varying patterns of shifted f0. R-R: raised precursor, raised following context; R-L: raised precursor, lowered following context; R-U: raised precursor, unshifted following context; L-R: lowered precursor, raised following context; L-L: lowered precursor, lowered following context; L-U: lowered precursor, unshifted following context; U-R: Unshifted precursor, raised following context; U-L: unshifted precursor, lowered following context; U-U: unshifted precursor, unshifted following context. More negative values indicate a higher proportion of low level responses, more positive values indicate a higher proportion of high level responses. Error bars indicate standard error.

As shown in Fig. 3, scores tended to increase when the context conditions changed from raised to unshifted to lowered, suggesting an increasing proportion of high level responses. In order to determine the relative effects of shifting the preceding versus following context, a two-way repeated measures ANOVA was calculated with three levels of preceding shift (raised, lowered, unshifted) versus three levels of following shift (raised, lowered, unshifted).

Results of the ANOVA showed that the effect of both the preceding and following contexts were statistically significant: Preceding, $F(2,22)=32.44$, $p<0.001$; Following, $F(2,22)=46.36$, $p<0.001$. On average, sentences with a raised preceding context were rated with a score of −3.4, as compared with 0.9 for lowered precursors and −0.1 for unshifted precursors. Similarly, sentences with a raised following context averaged −4.8, while lowered following contexts averaged 3.5, and unshifted following contexts averaged −1.3. The interaction between preceding and following contexts was not significant, $F(4,44)=2.08$, $p=0.10$.

Examination of means suggests that raising the following context had a stronger effect than raising the preceding context (raised following −4.8 versus raised preceding −3.4), while lowering the preceding and following contexts had similar effects (lowered following 3.5 versus lowered preceding 3.4). Because the experimental design means that cells with a raised precursor partially overlap with cells with a raised following context (e.g., the raised-raised sentences) these general patterns cannot be tested directly as part of the main ANOVA. However, it is possible to contrast the effects of raising and lowering the precursor and following contexts by comparing individual cells in the complete interaction. For example, when both the precursor and following context are raised, the mean response is most negative (−7.92), suggesting that listeners would most likely perceive the target word as having a low level tone in this condition. When neither context is shifted the mean score is virtually zero (−0.08), suggesting that listeners would most likely perceive the target word as having a midlevel tone in this condition. Finally, when both contexts are lowered, the mean response is the most positive (5.00), suggesting that listeners would most likely perceive the target word as having a high level tone in this condition. Raising the context increases the proportion of low level responses, while lowering it increases the proportion of high level responses, and the overall magnitude of the raising effect is somewhat larger than that of the lowering. Thus, these results correspond quite well to the findings of the previous experiment.

By examining specific pairs of context combinations using post-hoc (Tukey HSD, $\alpha=0.05$) analysis it is possible to tease apart the separate effects of raising and lowering the f0 of both preceding and following contexts. For example, when the precursor was raised and the following context was unshifted, the mean score was −4.08, but when the following context was raised and the precursor was unshifted the mean score was −4.00 (comparison of points *A* and *B* in Fig. 2), and this difference was not significant. In contrast, there

were significant differences between the effects of lowering the precursor as compared with the following context (0.25 precursor lowered versus 3.66 following context lowered; compared to the unshifted precursor and unshifted following context score of −0.08). These results suggest that lowering the precursor alone had little or no effect on listeners' responses (no different than the unshifted/unshifted context). In contrast, raising either the precursor or following context had a strong effect, as did lowering the following context.

In order to determine whether the precursor or following context had the stronger effect, the raised-lowered (RL) and the lowered-raised (LR) conditions were compared. If the precursor and following contexts had the same degree of effect on tone perception, then the responses to these two conditions should be identical. However, as seen in Fig. 3, responses to the RL condition were generally slightly positive (mean=1.75), indicating a greater prevalence of high level responses (consistent with the following context), while responses to the LR condition were somewhat more negative (mean=−2.42), indicating more low level responses (again, consistent with the following context). The difference between these two conditions was significant according to post-hoc (Tukey HSD) analysis.

A similar comparison can also be carried out by computing the magnitude of the effect of shifting only one of the two contexts while leaving the other context unshifted. For example, the difference between the raised following context (with unshifted precursor context, UR) and the lowered following context (with unshifted precursor context, UL) gives an effective magnitude of 7.67 for shifting the following context. In contrast, the magnitude of the effect of shifting the preceding context when the following context is held constant at the unshifted value was 4.33 (RU versus LU). Although there is not a significant difference between these two magnitudes (7.67 vs 4.33), $t(11)=1.98$, $p=0.07$, the overall pattern supports the hypothesis that the following context had a greater influence than the preceding context on normalization of Cantonese lexical tones.

C. Discussion

Results of experiment 2 suggested that both raising and lowering both the preceding and the following contexts had significant effects on Cantonese tone normalization. As in experiment 1, the effect of raising the f0 of the context one semitone was greater than that of lowering it by the same amount. Furthermore, the effect of changing the following context was larger than that of changing the preceding context. When the two contexts gave contradictory information, listeners appeared to depend more on the following context to perform tone normalization.

This pattern of results is broadly consistent with the findings of Lin and Wang (1985), in that they also found an effect of following context on the perception of lexical tone. In addition, other kinds of normalization, including phonetic context normalization (Johnson and Strange, 1982) and speaking rate normalization (Hirata and Lambacher, 2004; Newman and Sawusch, 1996) have been shown to be affected by subsequent as well as preceding extrinsic context.

Following these results, the Pitch Range Assessment Model may need to be extended to include some reference to the role of context location (in addition to the changes indicated by the results of experiment 1). Specifically, the results of experiment 2 suggest that listeners give more weight to the following context than to the preceding context, in a manner similar to other aspects of auditory pitch processing (Brady, *et al.*, 1961; Ciocca and Darwin, 1999). One might conceptualize such a process either as a moving window of analysis (e.g., a running average) or in terms of a feedback loop system. In a moving window type of system, the listener tracks the average f0 within a particular time frame (the window width), but the window moves through the signal over time. As the window moves away from earlier-occurring portions of the signal their importance declines, while the relative contribution of later-occurring information increases as the window moves toward/over them. Obviously, for post-target information to play a role in determining the target syllable's tone the later-occurring information must have already occurred by the time the decision is made. Thus, implicit in this model is the idea of waiting until some (unspecified) time after the target to make a final decision regarding the identity of that target [cf. discussion of speaking rate normalization by Miller and Dexter, (1988)].

Such a process is made more explicit in a feedback-loop based system. In a feedback-based system (cf. Nusbaum and Schwab, 1986) the listener is presumed to be continuously updating a hypothesis about the identity of the target on the basis of both internally generated and externally available information. Thus, the hypothesized identity of the target syllable could change as a consequence of later-occurring information.[6] On the basis of the present evidence, it is not possible to distinguish between the two models. Furthermore, it is not entirely certain that either model is necessarily involved in natural speech processing. Indeed, it is possible that listeners' preference for using later-occurring information to judge the identity of a medially presented target syllable may derive from the nature of the experimental task: Hear a sentence and report the identity of a given syllable. When carrying out this task it seems highly likely that listeners might (consciously or unconsciously) choose to reserve judgment until the entire sentence has been presented. Whether or not this kind of delayed processing takes place in more natural speech perception situations remains to be determined.

Finally, it is also possible that the preference for using later-occurring information in the present experiment provides support for the original Pitch Range Assessment Model, albeit in a more complex form. In the present experiment the following context had a larger pitch range than the preceding context. Tone number values ranged from 2 (low) to 5 (high) in the following context, but only from 2 to 3 (mid) in the preceding context. Thus, the following context provides a more complete picture of the pitch range of the speaker than does the preceding context. Thus, it is possible that listeners chose to reserve judgment about the identity of the target syllable because they knew that the following context would provide more information about the overall pitch range of the talker. Future experiments might control for this

possibility by providing equivalent tonal information in both preceding and following contexts. Still, all possible interpretations of these results suggest that, when possible, later-occurring information is incorporated into the tonal decision-making process, and is even preferred under certain task and stimulus conditions.

IV. EXPERIMENT 3

Research on phonetic context effects in speech perception frequently addresses the question of whether contextual effects derive from mechanisms specific to the perception of speech, or whether they result from more general auditory processes of assimilation and/or contrast that operate independently of whether the sounds being perceived are human speech or not (cf. Holt *et al.*, 2000). Similarly, it is possible that the talker-dependent normalization of lexical tones might result from the operation of a more general auditory process related to the perception of the pitch of complex tones or voices more generally, rather than one specifically related to the perception of linguistic tones. Wong (1998) showed that tone normalization could occur even across two languages, but in that case the listeners were familiar with both of the languages they heard, and both the context and target constituted meaningful utterances in one of the two languages. In order to determine whether normalization of lexical tones depends on listeners' ability to process the speech signal in a linguistically meaningful manner, the third experiment evaluated normalization of lexical tones presented within a context that could not be interpreted linguistically.

A. Methods

1. Subjects

Twelve college-aged, native Cantonese speakers (3 men and 9 women) with normal hearing and no history of speech or language disorder participated in this study. Note that these participants were recruited simultaneously with those in experiment 2, and assignment to either experiment 2 or experiment 3 was done on a gender-matched (but otherwise random) basis. None of these participants had taken part in any previous study on lexical tone normalization.

2. Stimuli

Stimuli were identical to those used in experiment 2, except that the context portions of each stimulus were rendered unintelligible by extracting the f0 contour and resynthesizing it using the "hummed" neutral vocal tract ([ə]) function in Praat 3.9.27.[7] As in experiment 2, there was a total of nine context conditions (3 levels of preceding context $\times 3$ levels of following context).

3. Procedure

Experimental procedures were identical to those used in experiment 2.

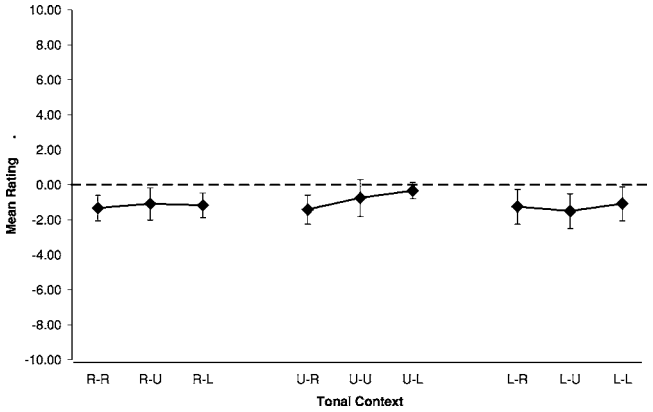

FIG. 4. Mean response scores (see the text) for tokens presented in unintelligible (/ə/) contexts with varying patterns of shifted f0. R-R: raised precursor, raised following context; R-L: raised precursor, lowered following context; R-U: raised precursor, unshifted following context; L-R: lowered precursor, raised following context; L-L: lowered precursor, lowered following context; L-U: lowered precursor, unshifted following context; U-R: unshifted precursor, raised following context; U-L: unshifted precursor, lowered following context; U-U: unshifted precursor, unshifted following context. More negative values indicate a higher proportion of low level responses, more positive values indicate a higher proportion of high level responses. Error bars indicate standard error.

B. Results

Responses were scored as in experiment 2, and results are shown in Fig. 4. As in experiment 2, a two-way repeated measures ANOVA was calculated with both direction of shift (raised, unshifted, lowered) and location of shift (preceding, following) as factors. In this case, results showed no significant effects of any kind. Listeners gave far more mid level responses (53.6%) than either low level (28.7%) or high level (17.7%) responses, suggesting that they were responding primarily to the absolute fundamental frequency of the target syllable (originally produced as a mid level tone) independently of the f0 of the surrounding context.

C. Discussion

In this experiment, listeners seemed unable or unwilling to use linguistically meaningless context for the purposes of tone normalization. It is possible that listeners treat nonword stimuli differently for the purposes of tone normalization, but previous studies in other languages have showed that listeners are able to make linguistically informed judgments about pitch patterns even when producing or perceiving nonsense words (e.g., Pierrehumbert, 1979). Thus, it may be assumed that listeners in the present study were able to perceive the pitch patterns of the context sentence, but for some reason failed to use the schwa-only context as a cue to the talker's fundamental frequency for the purposes of tone normalization. It is less clear whether listeners were consciously or unconsciously tuning out the schwa context (if such conscious control over a seemingly basic speech perceptual phenomenon is even possible). However, in either case, it seems plausible that some property of the context failed to match sufficiently with the corresponding property in the target syllable, allowing listeners to dissociate (consciously or unconsciously) the two types of signals and consider only the target syllable in making their tonal judgment. There are at least

three possible ways in which the context and target were mismatched, one linguistic and one nonlinguistic.

First, it is possible that listeners did not hear the [ə]-resynthesized stimuli as speech at all. While previous research suggests that nonspeech sounds are capable of supporting other kinds of (segmental) contextual normalization processes (Holt *et al.*, 2000), it is not clear that tone normalization necessarily operates according to identical principles. Further research is needed to determine whether or not manifestly nonspeech signals (e.g., pure tones) can function as a context for the purposes of tone normalization. Second, even if listeners did hear the hummed context as speech, the sound [ə] is not a phoneme in Cantonese, and therefore listeners may have ignored it because it is too "foreign" to be relevant for making tonal decisions. This is possible, but not likely, as Wong (1998) has already shown that a familiar foreign language context (English) can provide sufficient pitch information to support at least some degree of lexical tone normalization. Finally, it is possible that listeners treated the schwa-context as if it had been produced by a different talker. In the present stimuli, no attempt was made to match any talker identity-related parameters of the context to that of the target. Thus, it is possible that lexical tone normalization depends on the perception that both the context and the target were produced by the same talker.

V. EXPERIMENT 4

The process of talker normalization represents an interface between processes of talker identification and recognition on the one hand, and processes of speech perception on the other. However, there has been relatively little research on the degree to which the perception of talker *identity* influences talker normalization.[8] In some studies of tone normalization, all tokens were produced by the same talker (e.g., Wong, 1998; Wong and Diehl, 2003). In those cases in which talker and target were not produced by the same talker [e.g., Moore and Jongman (1997); Leather (1983)], synthetic stimuli were presented within a sentential context produced by a natural talker. However, both of these studies attempted to synthesize the target tokens such that there was little possibility that listeners might identify the target as having been produced by a different talker than the context. If tone normalization depends on the operation of mechanisms for talker identification, then we would expect a perceptible difference in talker between context and target to disrupt it. If hearing a target and context produced by a different talker disrupts the process of tone normalization even when the context consists of linguistically meaningful, natural speech in the listeners' native language, this would provide strong support for the hypothesis that tone normalization (and, by extension perhaps talker normalization in general) are closely related to processes of talker recognition or identification. On the other hand, if tone normalization occurs across obviously different talkers, this might suggest that perceptual normalization, at least for tone, may rely on the operation of more general auditory mechanisms that are not specific to the perception of speech *per se*.

A. Methods

1. Subjects

Twenty-two native Cantonese speakers (16 women, 6 men, aged 21–24 years) with normal hearing volunteered to participate in this experiment. All participants were students in the Division of Speech and Hearing Sciences at the University of Hong Kong.

2. Stimuli

Two native Cantonese men who were judged by P.C.Y.C. (a final-year student in speech and language pathology who had received phonetic training) as having similar pitch ranges, but clearly different-sounding voices, were selected as speakers. They read the same semantically neutral context sentence with embedded mid level tone and stimuli were recorded using the same procedures as in experiment 2. The average f0 for the carrier phrase with the mid level tone was 123.9 Hz for speaker 1 and 112.3 Hz for speaker 2. One production of the target sentence with the target tone for each speaker was selected by two final-year speech and language pathology students (one of whom was P.C.Y.C.). The two sentences were selected so that the target word could clearly be identified as a mid level tone, and both sentences had a perceptually similar rhythm. The three segments of the carrier sentence (preceding context, target syllable, and following context) were measured for both speakers to ensure that they had similar durations across the two speakers: the duration of the preceding context was about 710 ms for speaker 1 and 695 ms for speaker 2; the following context was about 720 ms for speaker 1 and 745 ms for speaker 2. The duration of the target word was set to 250 ms (close to the duration of the natural token produced by speaker 2) for both speakers using the manipulation functions of the Praat software. Measurements of f0 for each word for each sentence were made, and speaker 1 was identified as having a larger overall range of f0 values across the sentence, and was designated as the model talker.

After this, the amplitude of each of the two sentences (one per speaker) was peak-normalized, and the fundamental frequency pattern of each word within each sentence was resynthesized using the PSOLA algorithm of the Praat software. For resynthesis, the f0 values of the two sentences were first extracted using the default autocorrelation algorithm in Praat. Speaker 2's f0 values for each word within each sentence were then manually adjusted to new f0 values that were close to the f0 values for that word of speaker 1 (the model talker). Note that, although the average f0 of the resynthesized target (106 Hz) was more similar to the average context f0 of speaker 2, this was because the range of speaker 1 was larger. Therefore, the words /teng55/ and /bei25/ ended at a higher f0 for speaker 1 than speaker 2, which is why the average f0 of the sentence context was higher for speaker 1. Exact F0 values were set such that the sentences retained a natural-sounding intonation and the target word had correct tonal identity as judged by P.C.Y.C. and two other native Cantonese speakers. The carriers for both speakers were then resynthesized with the new, similar F0 values using the PSOLA algorithm. After resynthesis, the F0

TABLE II. Fundamental frequencies of stimuli after modification (experiment 3). Note: Entries in italic indicate stimuli presented to the same talker group. Entries in Roman indicate stimuli presented to the different talker group.

	Fundamental frequency (Hz)		
Context	Carrier phrase	Target word talker 1	Target word talker 2
Talker 1: raised	97–133	*106*	106
Talker 1: unshifted	92–125	*106*	106
Talker 1: lowered	86–118	*106*	106
Talker 2: raised	96–133	106	*106*
Talker 2: unshifted	91–125	106	*106*
Talker 2: lowered	86–117	106	*106*

values of the sentences for the two speakers were determined to be very closely matched (within a few Hertz for each word): The average f0 of the target words in the resynthesized stimuli was 106 Hz for both speakers.

After generating these sentences of matched duration and f0 characteristics, two additional sentences were created by splicing the target word from speaker 1 into the carrier phrase spoken by speaker 2, and vice versa, thereby creating two additional stimulus sentences in which the target syllable was produced by a different talker than was the context in which it appeared. Thus, there were four base sentences differing only in terms of the (nonpitch) vocal properties: Two with matching talkers for context and target, and two with mismatched talkers for context and target. The f0 values of the context (nontarget) portion of each of these four base sentences were then either raised, unshifted or lowered by one semitone as in experiment 1. F0 properties for the resulting twelve stimuli are shown in Table II.

3. Procedure

Participants were randomly assigned to one of two groups. In the same talker group, listeners heard only sentences in which the same talker produced the target syllable and the context in which it was embedded. Listeners in the different talker group heard only stimuli in which the target syllable and surrounding context were produced by different talkers. Experimental procedures were otherwise identical to those of Experiment 2.

B. Results

As in the first experiment, the proportion of expected responses was calculated for each condition. A three-way, mixed factorial ANOVA was calculated, with the type of carrier phrase (Same Talker versus Different Talker) as a between groups factor and the direction of f0 shift (raised, lowered, unshifted) and target talker (talker 1, talker 2) as within groups factors. Results of the ANOVA indicate no significant main effects of type, $F(1,20)=1.18$, $p=0.29$, or talker, $F(1,20)=0.88$, $p=0.36$. There was a significant effect of shift, $F(2,40)=7.20$, $p=0.002$, such that lowered contexts resulted in 53% expected (high level) responses, unshifted contexts result in 80% expected (mid level) responses, and raised contexts result in 70% expected (low level) responses.

Significant two-way interactions were found between talker and type, $F(1,20)=11.43$, $p=0.003$, shift and type, $F(2,40)=5.17$, $p=0.01$, and talker and shift, $F(2,40)=10.15$, $p<0.001$.

These interactions can be explained by exploring the significant three-way interaction between talker, shift, and type, $F(2,40)=6.55$, $p=0.003$, as shown in Figs. 5(a) and 5(b). Post-hoc (Tukey HSD, $\alpha=0.05$) analysis revealed that the primary sources of the observed interactions were (1) a significantly lower proportion of expected responses in the same talker, lowered context conditions (leftmost circles in each graph, and significantly lower than either the raised or unshifted same talker conditions), and (2) a significantly lower proportion of expected responses in the raised, different talker condition with the target produced by talker 2 (rightmost closed squares in Fig. 5(a), and significantly lower than for the corresponding condition of tokens produced by talker 1 but presented within a talker 2 context).

VI. GENERAL DISCUSSION

The results of the same talker condition (T1/T1 and T2/T2) in the present experiment were comparable to those found in experiments 1 and 2, such that shifting the f0 of the context upward resulted in more low level responses, while shifting it downward resulted in more high level responses and the effect of the downward shift was not as strong as the upward shift. However, the pattern of responses in the different talker condition provides additional insight into the process of tone normalization. First, we observed that, in the different talker condition (T1/T2 and T2/T1), listeners were overall more inclined to respond as expected, and this appears to be primarily the result of an increase in the effectiveness of the lowered shift condition in the different group as compared with the same talker group: Both the T1/T2 and T2/T1 stimuli showed a greater proportion of expected responses with a lowered context than did either the T1/T1 or T2/T2 conditions in the same context. Second, although post hoc (Tukey HSD, $\alpha p=0.05$) tests revealed no difference between shift conditions when averaging across talkers, in the raised condition there is evidence of an effect of individual talkers. When the talkers who produced the target and context did not match (squares in both graphs), in the raised condition there was a significant drop in expected responses to talker 2 targets in talker 1 context as compared with talker 1 targets in talker 2 context. This was the only significantly different comparison between the two talkers. Thus, mismatching the talker of the target and the context seems to improve listeners' normalization of lexical tones in the lowered context condition, but in the raised condition normalization is disrupted when talker 1 targets are placed in a talker 2 context (but not vice versa).

The disruption of normalization when talker 1 targets are presented in a talker 2 context could be explained by differences in the sound of the speech of the two speakers. Although the f0 of the two speakers in the stimuli was adjusted to be identical, the rest of the speech [the perceived

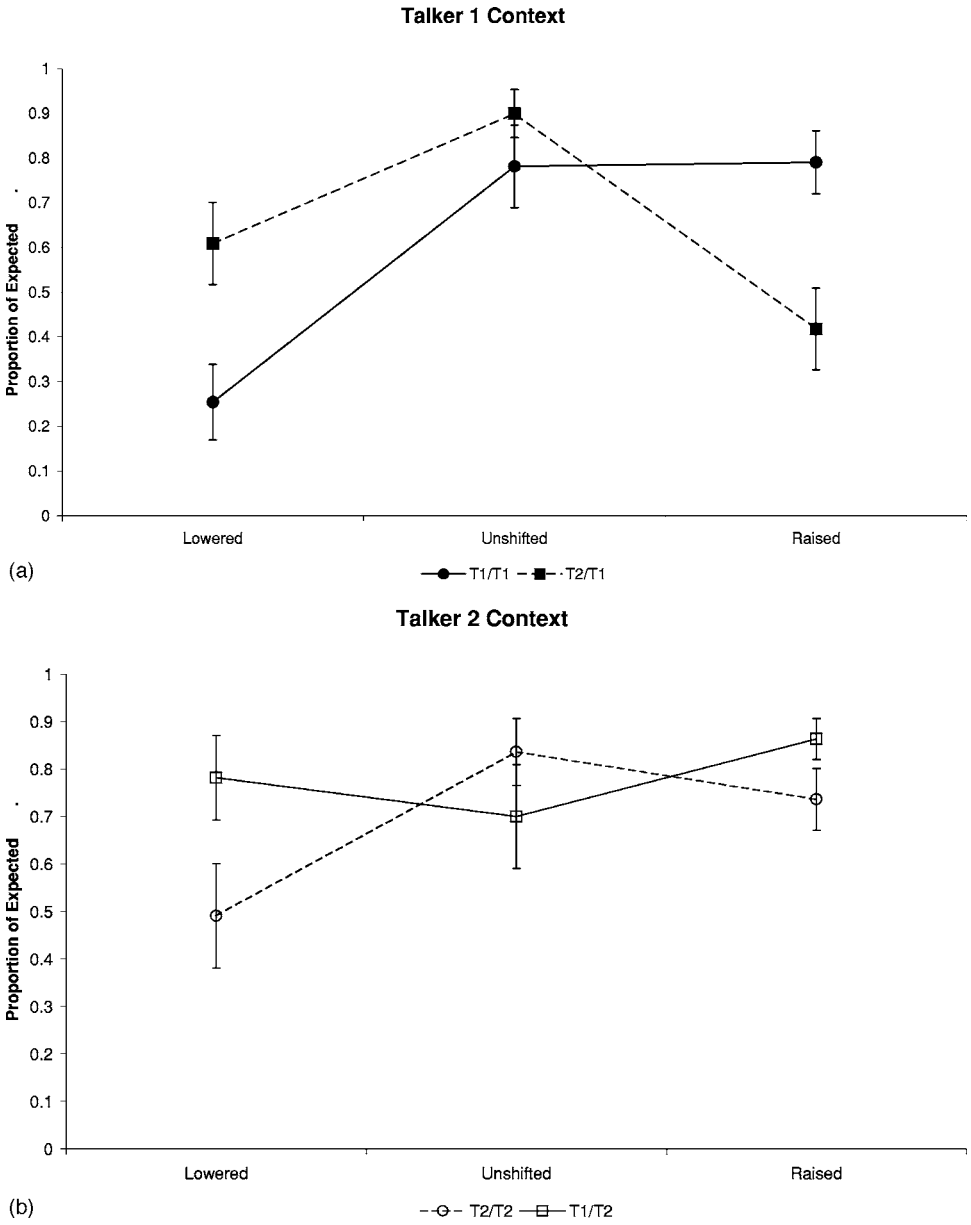

FIG. 5. (a) Mean proportion of expected responses for targets produced by talker 1 (T1/T1, circles, solid line) or talker 2 (T2/T1, squares, solid line) embedded within context sentences produced by talker 1. Labels first list the talker who produced the target, followed by the talker who produced the context. In (a) and (b), stimuli with targets produced by talker 1 are indicated with solid lines, while stimuli with targets produced by talker 2 are indicated with dashed lines. Stimuli with contexts produced by talker 1 are indicated with closed symbols, while contexts produced by talker 2 are indicated by open symbols. Stimuli where the same talker produced both target and context are indicated with circles, while those in which the talker of the target and context differ are indicated with squares. Error bars indicate standard error. (b) Mean proportion of expected responses for targets produced by talker 1 (T1/T2) or talker 2 (T2/T2) embedded within sentences produced by talker 2. Nomenclature and symbols as in (a). Error bars indicate standard error.

"timbre," including properties related to both laryngeal (voice quality, source spectrum) and supralaryngeal differences (vowel formant spacing, relative timing of articulators)] of each talker was (intentionally) noticeably different. Singh and Hirsh (1992) found that variation in timbre could be perceived as a change in pitch when stronger pitch cues such as the fundamental frequency were kept constant. In the present experiment, ten additional listeners exposed to the same talker stimuli rated talker 2 as having a higher pitched voice than talker 1 even though the f0 of the two talkers' stimuli was identical. Since the timbre of a speech signal depends on both the formant (filter) properties and the harmonic (source) properties of the talker's voice, and the source properties were identical, this suggests that differences in the two talkers' individual formant patterns must contribute to the difference in their perceived pitch. The source/filter resynthesis algorithm used for shifting f0 in the present experiment does not alter formant frequencies. Thus, each talker's filter function remained constant across shift conditions. That is, talker 1's speech must have been perceived as having an overall lower pitch than talker 2's speech in all conditions. Therefore, when the target syllable produced by talker 2 was placed within the raised context produced by talker 1, the perceptual magnitude of the f0 shift effect was reduced because of the higher pitch perceived for talker 2 as compared with talker 1.

This effect of timbre, however, cannot explain the overall superior effect of the lowered shift in the different talker condition as compared to the lowered shift in the same talker condition. Indeed, post-hoc (Tukey HSD, $\alpha=0.05$) analysis showed that listeners in the lowered shift condition who heard talker 1 targets presented in talker 2 context showed a greater proportion of expected (high level) responses ($p=0.05$) compared with those who heard talker 1 targets in talker 1 context, despite the fact that timbre differences should have caused a smaller perceived effect of shift in the T1/T2 condition because the talker 2 context should have sounded less lowered than the corresponding talker 1 context

in the same (lowered) shift condition. Conversely, we would have expected timbre differences to produce an enhancement of the shift effect in the raised condition when presenting talker 2 targets in talker 1 context (T2/T1) as compared to the T2/T2 presentation because the perceived pitch of the talker 2 target should have been lower, or the perceived pitch of the T1 context should have been higher, than the f0 measurements alone would have indicated, or both. However, there was no significant difference in proportion of expected responses to talker 2 targets in talker 1 versus talker 2 contexts.

The specific details of how timbre affects the perception of pitch for the purposes of normalization of tones remains to be explained, but the results of the lowered condition clearly show that normalization can actually be stronger when target and context do *not* match (as compared to when they do), supporting the hypothesis that tone normalization does not *require* the perceived continuity of talker.

If perceptual normalization of lexical tone does not require a perceived continuity of talker, why, then, did listeners fail to use pitch information from extrinsic context in experiment 3, where the words in the context had been replaced by [ə]? At this point, it seems reasonable to conclude that listeners were able to recognize the series of schwa-syllables as being meaningless, and were therefore able to treat them as irrelevant to tonal processing. This suggests that the process of tone normalization is, in some sense, also a linguistic one, and does not depend solely on the automatic operation of nonlinguistic, auditory processes. While the determination of pitch for the purposes of talker normalization appears to function independently of perceived talker continuity (and therefore may be the result of a more general auditory process), the application of such information to tone normalization may require that listeners recognize the context as being linguistically meaningful. However, these conclusions are only tentative, pending the results of further research on the ability of listeners to normalize lexical tones in the context of a talker speaking a real language, but one that is unknown to the listener [though see Jongman and Moore (2000) for preliminary results in this direction]. If future research demonstrates that even appropriately constructed nonspeech contexts can induce tone normalization effects, this would instead suggest that the phenomenon of tone normalization, and perhaps talker normalization in general, might derive from more basic auditory processes for maintaining perceptual constancy across contexts (see Holt, 2005 for similar discussion).

The application of general processes of perceptual constancy to the specific task of talker normalization might be universally successful under normal (real world) task conditions, and yet could still perform less than optimally in specific (typically laboratory-related) contexts. For example, in the real world it is relatively improbable that talker identity would change fluently in the middle of a sentence, and even if it did there would necessarily be some additional cues to the change including changes in the spatial location of the voice, as well as the strong possibility of additional differences in intrinsic vocal properties such as timbre and fundamental frequency. However, in the laboratory it is possible to artificially optimize the transition between context and target in order to preserve most cues to continuity (e.g., by matching fundamental frequencies and maintaining the same spatial location for the voice, as in the present experiment). Under such artificial circumstances listeners may be "fooled" into applying a valid, general perceptual strategy across talkers in the same way that many visual illusions result from the correct operation of visual strategies that are adequate for performance in the real world but yield predictably illusory percepts when "tricked" in particular ways in the laboratory (Marr, 1982, p. 30).

VII. SUMMARY

In order to accurately identify the tonal category of a given syllable, tone language speakers use pitch information obtained from extrinsic speech to estimate the expected pitch properties of possible tones (the talker's tone space). The perceived lexical tone of a syllable is based on a match between the perceived pitch of the syllable and the representation of that tone within the mental tone space induced for that talker on the basis of pitch information available from context. The results of the experiments presented here provide further information relevant to evaluating and modifying the Pitch Range Assessment Model. The original formulation of the PRAM proposed that listeners' estimate of a given talker's pitch space depends on the pitch range present in the extrinsic context. Results from the present experiment 1 suggest instead that the mechanism for estimating a given talker's tonal space involves a process of extrapolation from the talker's average f0. Previous studies of tone normalization examined the effects of either preceding context (e.g., Fox and Qi, 1990; Leather, 1983; Moore and Jongman, 1997; Wong, 1998; Wong and Diehl, 2003) or of a single post-target syllable (Lin and Wang, 1985) but did not specifically compare the two types of conditions to determine which (if either) had a greater effect on normalization. Furthermore, no previous study has specifically attempted to determine whether tone normalization is affected by perceived differences in talker. Results of the present experiments 2 and 4 suggest that pitch information for tone normalization appears to be integrated both across different regions of context (both preceding and following the target) as well as possibly across different sources of pitch perception (f0 and timbre), although the evidence supporting a role for timbre-based pitch information is only suggestive at this point. Finally, the results of the present experiments 3 and 4 suggest that the process of tone space estimation does not function in a talker-specific manner, but may still depend on the listeners' linguistic knowledge. That is, tone normalization does not seem to be dependent upon listeners' perception that the entire utterance was produced by a single talker, but it does appear to be necessary that the listener be able to recognize the context as meaningful speech. Whether or not the speech must be meaningful to listeners is still undetermined [though see Jongman and Moore (2000) for evidence that English listeners do show some perceptual normalization of Mandarin tones, albeit not the same pattern as Mandarin listeners].

Further research on the possible effects of unfamiliar linguistic contexts will be necessary to investigate this hypothesis more thoroughly.

ACKNOWLEDGMENTS

We would like to thank Yukari Hirata, Phil Rose, and two anonymous reviewers for helpful comments given on an earlier draft of this manuscript. The work described in this paper was funded by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. HKU 7913/00H).

[1]In order to compare tones across languages, tones in this paper are transcribed in a manner similar to the Chao tone number system used by Bauer and Benedict (1997). This system reflects the relative pitch of the syllable at onset and offset within a 5-point scale from the bottom (1) to top (5) of the talkers' normal pitch range. Thus, for the Cantonese tonal system 55 =high level, 25=high rising, 33=mid level, 21=low falling, 23=low rising, and 22=low level. For Mandarin, 55=tone 1 (high level), 35=tone 2 (rising), 214=tone 3 (dipping), and 51=tone 4 (Falling).

[2]Moore and Jongman (1997) actually used a continuum of syllables ranging in both frequency and time of the turning point of f0 from a good 25 (tone 2) to 214 (tone 3) syllable. The description of the effect they identified is given here with respect to a single syllable in the middle of the continuum in order to emphasize the role of contextual differences.

[3]Note that Wong (1998) was able to increase the effectiveness of the English context until it matched that of the Cantonese context by increasing the degree of difference between the average f0 of the context and that of the target. Cantonese listeners required a four semitone upward shift or a three semitone downward shift of the English context before they reported that the target had a low level or high level tone (respectively) to the same degree that they did with a three semitone upward or two semitone downward shift in a Cantonese precursor. These results not only suggest that listeners' response patterns are gradient (as confirmed in the present experiment 2), but also further support the hypothesis that listeners perform tone normalization differently depending on the language of the context.

[4]With respect to translation, note that the characters /ji33/意(meaning) and /ji22/ (two) are commonly produced as single words and thus translating them as such poses no problem. In contrast, the status of /ji55/醫is not as clear. This character does not typically appear outside of multisyllabic collocations, but its most common use is in the two-word combination that means *doctor* (followed by生/sa:ŋ55/). Thus, the closest English interpretation of the semantic contribution of /ji55/醫to this collocation is *doctor*. With respect to asking listeners to respond to single characters, native speakers have no trouble pronouncing the character醫/ji55/ in isolation. Moreover, in the present study listeners were instructed to respond on the basis of the *character* they heard. Thus, although the character 醫/ji55/ does not typically appear in isolation, the task of identifying this item was not markedly more difficult or unusual than identifying either of the other two tokens.

[5]Nonexpected responses in the raised and lowered conditions were generally 33 (mid level) responses, reflecting either (or both) the fact that the f0 of the target syllable was in fact originally that of a mid level tone, and/or that it is very unlikely to misidentify a high level tone as a low level tone, or vice versa. Thus, a mid level response was the most likely default response both on the basis of its intrinsic f0 properties, and simply because the midlevel tone lies between the high and low level tones.

[6]This process is also compatible with the theory of Analysis by Synthesis (e.g., Stevens and Halle, 1964).

[7]The Praat resynthesis menu selection reads "To Sound (hum)" but the actual result is much closer to a schwa than a hum. The resulting sound is clearly non-nasal, and has formant frequencies of F1=600, F2=1400, F3 =2392, and F4=3412, quite similar to those measured by Peterson and Barney (1952) for the central unrounded vowel of American English.

[8]Although see work by Nygaard and colleagues (Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994) for important examples of work in this direction. This work does not, however, deal with the perception of lexical tones.

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin).

Boersma, P., and Weenink, D. (2001). "Praat: A system for doing phonetics by computer." Retrieved May 22, 2001, from University of Amsterdam, Institute of Phonetics Sciences Web site: http://www.praat.org (confirmed 15 April, 2005).

Brady, P. T., House, A. S., and Stevens, K. N. (**1961**). "Perception of sounds characterized by a rapidly changing resonant frequency," J. Acoust. Soc. Am. **33**, 1357–1362.

Chao, Y. R. (**1947**). *Cantonese Primer* (Harvard University Press, Cambridge, MA).

Ciocca, V., and Darwin, C. J. (**1999**). "The integration of nonsimultaneous frequency components into a single virtual pitch," J. Acoust. Soc. Am. **105**, 2421–2430.

Fok Chan, Y. Y. (1974). "A perceptual study of tones in Cantonese," Centre of Asian Studies, University of Hong Kong, Hong Kong.

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

Francis, A. L., Ciocca, V. C., and Ng, B.K. C., **2003**. On the (non)categorical perception of lexical tones," Perception & Psychophysics, **65**(6), 1029–1044.

Fu, Q.-J., Zeng, F.-G., Shannon, R. V., and Soli, S. D. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.

Hirata, Y., and Lambacher, S. G. (**2004**). "Role of word-external contexts in native speakers' identification of vowel length in Japanese," Phonetica **61**, 177–200.

Holt, L. L. (**2005**). "Temporally non-adjacent non-linguistic sounds affect speech categorization," Psychol. Sci. **16**, 305–312.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2000**). "Neighboring spectral content influences vowel identification," J. Acoust. Soc. Am. **108**, 710–722.

Johnson, K. (**1991**). "Differential effects of speaker and vowel variability on fricative perception," Lang Speech **34**, 265–279.

Johnson, K. (**1997**). "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 145–165.

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, T. L., and Strange, W. (**1982**). "Perceptual constancy of vowels in rapid speech," J. Acoust. Soc. Am. **72**, 1761–1770.

Jongman, A., and Moore, C. (**2000**). "The role of language experience in speaker and rate normalization processes," Proceedings of the Sixth International Conference on Spoken Language Processing, I, pp. 62–65.

Joos, M. A. (**1948**). "Acoustic Phonetics," Language **24**, 1–136.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

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.

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431–461.

Lin, T., and Wang, W. Y. (**1985**). "Tone perception," Chinese Linguistics J. **2**, 59–69 [in Chinese].

Marr, D. (**1982**). *Vision* (Freeman, New York).

Matthews, S., and Yip, V. (**1994**). *Cantonese: A Comprehensive Grammar* (Routledge, London).

Miller, J. L., and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psychol. Hum. Percept. Perform. **14**, 369–378.

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

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Newman, R. S., and Sawusch, J. R. (**1996**). "Perceptual normalization for speaking rate: Effects of temporal distance," Percept. Psychophys. **58**, 540–560.

Nusbaum, H., and Magnuson, J. (**1997**). "Talker normalization: Phonetic

constancy as a cognitive process," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 109–132.
Nusbaum, H. C., and Schwab, E. C. (**1986**). "The role of attention and active processing in speech perception," in *Speech Perception*, edited by E. C. Schwab and H. C. Nusbaum, Pattern Recognition by Human and Machines, Vol. **1** (Academic, San Diego), pp. 113–157.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker contingent process," Psychol. Sci. **5**, 42–46.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pierrehumbert, J. A. (**1979**). "The perception of fundamental frequency declination," J. Acoust. Soc. Am. **66**, 363–369.
Rose, P. (**2000**). "Hong Kong Cantonese citation tone acoustics—A linguistic-tonetic study," in Proceedings of the Eighth Australian International Conference on Speech Science and Technology, edited by S. Barlow, Australian Speech Science and Technology Association, pp. 198–203.
Rose, P., July 18, 2005 (personal communication).
Singh, P. G., and Hirsh, I. J. (**1992**). "Influence of spectral locus and F0 changes on the pitch and timbre of complex tones," J. Acoust. Soc. Am. **92**, 2650–2661.
Stevens, K. N., and Halle, M. (**1964**). "Remarks on analysis by synthesis and distinctive features," in *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form*, edited by W. Wathen-Dunn (MIT, Cambridge), pp. 88–102.
Vance, T. J. (**1976**). "An experimental investigation of tone and intonation in Cantonese," Phonetica **33**, 368–392.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **60**, 198–212.
Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**, 25–47.
Wong, P. C. M. (1998). "Speaker normalization in the perception of Cantonese level tones," Master's thesis, University of Texas at Austin (unpublished).
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.