General perceptual contributions to lexical tone normalization

Jingyuan Huang[a] and Lori L. Holt
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213

(Received 3 November 2008; revised 1 April 2009; accepted 5 April 2009)

Within tone languages that use pitch variations to contrast meaning, large variability exists in the pitches produced by different speakers. Context-dependent perception may help to resolve this perceptual challenge. However, whether speakers rely on context in contour tone perception is unclear; previous studies have produced inconsistent results. The present study aimed to provide an unambiguous test of the effect of context on contour lexical tone perception and to explore its underlying mechanisms. In three experiments, Mandarin listeners' perception of Mandarin first and second (high-level and mid-rising) tones was investigated with preceding speech and non-speech contexts. Results indicate that the mean fundamental frequency (*f*0) of a preceding sentence affects perception of contour lexical tones and the effect is contrastive. Following a sentence with a higher-frequency mean *f*0, the following syllable is more likely to be perceived as a lower frequency lexical tone and vice versa. Moreover, non-speech precursors modeling the mean spectrum of *f*0 also elicit this effect, suggesting general perceptual processing rather than articulatory-based or speaker-identity-driven mechanisms. © *2009 Acoustical Society of America.*
[DOI: 10.1121/1.3125342]

PACS number(s): 43.71.An, 43.66.Ba, 43.71.Hw [RSN] Pages: 3983–3994

I. INTRODUCTION

A. Speaker normalization effects in phonetic categorization

The acoustics of speech are notoriously variable across speakers. Some of this variability is the result of anatomical and physiological differences in the instrument of speech production, such as larger (and differently-proportioned) vocal tracts of male vs female speakers. Other variability such as foreign accent and dialect stems from linguistic and sociolinguistic experience. A result of all this variability is that phonetic categories and decision bounds founded on experience across a variety of talkers may produce miscategorization in application to any *particular* talker. Thus, it has long been suggested that speech categories must be tuned dynamically to the speech of the current talker, for example, by shifting the representation of the individual sounds or by influencing the relevant phonetic category space to which sounds are mapped (Ladefoged and Broadbent 1957). Within the field of speech perception, the accommodation of talker-specific characteristics is referred to as "talker" or "speaker" normalization (e.g., Johnson and Mullennix, 1997).

One of the most influential experiments testing speaker normalization comes from Ladefoged and Broadbent (1957), who demonstrated that manipulating the voice in which a precursor sentence is spoken has a major effect on how listeners categorize a following vowel. Using speech synthesis, Ladefoged and Broadbent (1957) manipulated the frequencies of the first two formants of the sentence "Please say what this word is," resulting in six sentences that sounded like they were spoken by different talkers. Following these context sentences, participants heard synthesized "bVt" target syllables varying in their first two formant frequencies and approximating "bit," "bet," "bat," or "but." Listeners' vowel categorization was influenced by the characteristics of the preceding sentence context. For example, a vowel that was identified as "bit" by 88% of participants in context of the original sentence was identified as "bet" by 90% of participants when the sentence was manipulated to have a lower F1 frequency. In all, these results suggest that extrinsic context interacts with the intrinsic acoustic properties of a speech segment to tune how listeners categorize acoustic speech signals.

Ladefoged and Broadbent (1957) proposed that listeners recover talkers' vocal tract dynamics from context sentences and use this information to scale speech perception, mapping phonetic information available from the vowels of the context sentence onto a F1 × F2 space, and using this information to identify the target vowels by their relative position in this space. Further, Ladefoged and Broadbent (1957) proposed that this re-mapping be considered a speech-specific calibration, "best understood by the reference to the articulatory process in speech" (p. 103). In other words, listeners extract a speaker's vocal tract information from context and normalize perception according to the perceived vocal tract.

A series of subsequent experiments by Watkins and Makin (1994, 1996) calls into question the necessity of a speech-specific interpretation. Watkins and Makin (1994, 1996) conducted several variants of the Ladefoged and Broadbent (1957) task, substituting the "Please say what this word is" with noise analogs and context sentences played backwards. Critically, the effect of context persisted even when contexts did not preserve vocal tract or articulatory information, suggesting that the link between articulatory in-

[a]Author to whom correspondence should be addressed. Electronic mail: jingyuan@andrew.cmu.edu

formation and speaker normalization effects, as posited by Ladefoged and Broadbent (1957), may be tenuous.

One specific alternative interpretation for the shifts in speech categorization described as speaker normalization is that listeners tune perception according to the general distributional characteristics of preceding acoustic context. To make this concrete, it is helpful to consider again the effect reported by Ladefoged and Broadbent (1957). In general, productions of the vowels /I/ and /ɛ/ differ in F1 frequency with /I/ having a lower-frequency F1 (Peterson and Barney, 1952). Presumably, listeners have formed categories for these vowels based on regularities across speakers (or at least the typical values that define the contrasts in relation to other vowels). As a result, when a vowel is encountered with a low F1 it is categorized more often as /I/ as in "bit." However, the actual value that corresponds to a "low F1" appears to be relative to speech produced by a particular talker. As Ladefoged and Broadbent (1957) demonstrated, when the range of F1 in the context sentence is lowered, the same F1 value encourages /ɛ/ as in "bet" categorization. Considering the results of Watkin and Makins (1994, 1996), it is possible that the spectral change in mean frequency of F1, rather than its influence on perceived speaker identity or the articulatory information it may convey, is responsible for the observed effects on speech categorization. If so, effects of speaker normalization potentially may arise from general perceptual processes.

Some recent results investigating the influence of distributions of spectral energy on phonetic categorization support this possibility. Holt (2005) found that sentence-length sequences of non-speech sine-wave tones have a strong influence on categorization of subsequent speech. In these experiments, the sine-wave tone sequences sample a region of spectral energy defined with a particular mean acoustic frequency and variability about this mean; stimuli vary on every trial but sample a consistent region of the spectral space and respect the distributional characteristics that define a condition. Therefore, idiosyncracies of acoustic sampling cannot drive any observed effects of context. For these non-speech "acoustic histories" to impact speech categorization, the long-term distribution characteristics of the acoustic spectra must play a role. Much like the increments and decrements to F1 frequency that Ladefoged and Broadbent (1957) used to manipulate perceived talker in their synthesized context sentences, Holt (2005) made conditions in which the non-speech tone distributions sampled higher- and lower-frequency spectral regions. The resulting stimuli sound something like a higher- and lower-frequency melodies followed by a speech token. Listeners simply identified the speech sound. Holt (2005, 2006a, 2006b) found that the mean frequency of the preceding spectral distribution of non-speech tones exerts a strong influence on speech categorization.

Intriguingly, the speech categorization shifts produced by these simple sine-wave tone sequences mirror the directionality of the influence of acoustic manipulations to sentence contexts and their corresponding influence on speech categorization reported by Ladefoged and Broadbent (1957; see also Watkins and Makin, 1994). Specifically, they are contrastive. A lower frequency context, whether produced by decrementing the mean F1 frequency of a sentence or by shifting the mean frequency of a distribution of non-speech sine-wave tones to lower frequencies, influences listeners' speech categorization by pushing it toward higher-frequency alternatives [e.g., from /I/ to /ɛ/ in the case of the Ladefoged and Broadbent (1957) results]. Despite the commonality in their influence on speech categorization, it is important to highlight the differences in the information available in these two sets of contexts. Unlike the sentence contexts of Ladefoged and Broadbent (1957), the acoustic histories of Holt (2005, 2006a, 2006b) were composed entirely of non-speech stimuli, providing no information about vocal tract characteristics of the speaker, no sampling of the English phonetic space, and no reference to a human voice whatsoever. Nevertheless, they strongly influenced speech categorization. Listeners thus have proven to be very sensitive to the longer-term characteristics of the acoustic signal and adjust perception of speech in relation to statistical regularity in prior acoustic input, even when that input is non-speech. Whatever the mechanisms involved (see Holt, 2006b for speculation on mechanism), they must be rather broadly operative.

The prospect that general auditory processes not specific to speech and not requiring information about articulatory gestures or human vocal tracts may account for effects described as speaker normalization invites the possibility that the commonalities among the effects observed thus far might be applied to make predictions for other normalization challenges in speech perception. From these prior experiments (Holt, 2005, 2006a, 2006b), we would predict spectrally (or temporally, Wade and Holt, 2005) contrastive effects. In addition, they may be elicited by non-speech, as well as speech, precursors. Moreover, given the proposed generality of the auditory processing involved, we should expect effects to generalize beyond English to be present among native listeners of other languages.

In the present work, we exploit the perceptual challenges of lexical tone normalization in Mandarin Chinese listeners' perception to investigate these predictions. We first review the literature of lexical tone normalization in light of the possibility that contrastive general mechanisms may play a role. We then present three experiments to test the predictions outlined above.

B. Speaker normalization in lexical tones

Tone languages use pitch to contrast meaning. For example, Mandarin Chinese has four different lexical tones: high-level tone (tone 1), mid-rising tone (tone 2), low-falling-rising tone (tone 3), and high-falling tone (tone 4) (Ladefoged and Maddieson, 1996). As can be seen in Fig. 1, the *f*0 of Mandarin words changes mean *f*0 frequency (height) and contour (change in frequency) to shift word identity. The *f*0 trajectories are plotted in Fig. 1 for a single isolated syllable spoken by one talker. The large *f*0 differences across lexical tones in these circumstances belie the variability present in more natural utterances. In fact, the exact nature of the *f*0 characteristics of Mandarin words is highly variable across utterances and speakers. Thus, some

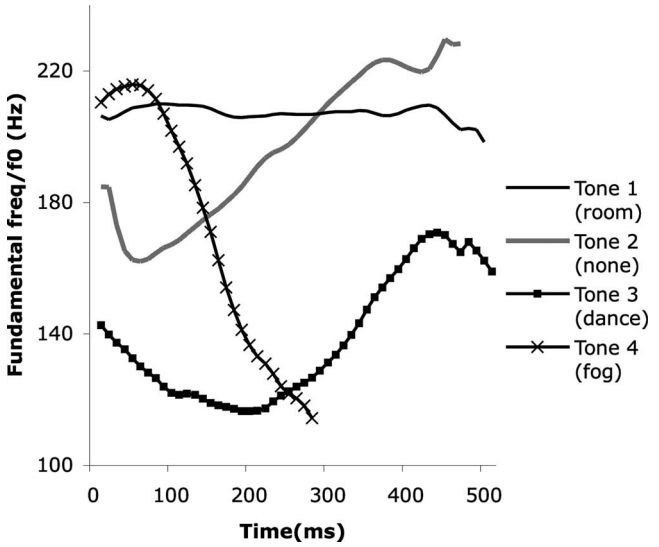

FIG. 1. *F*0 contour of Mandarin tones in the isolated syllable /wu/, measured from the speaker recorded to create stimuli for the present experiments. Changes in tone change the meaning of the syllable, as indicated by the English translations in the figure legend.

of the same perceptual challenges exist for distinguishing lexical tones across speakers and contexts that exist for phonetic categorization across talkers. For example, a lexical tone with a low frequency (like tone 3) that is produced by a higher-frequency voice might have an *f*0 similar to a higher-frequency lexical tone (like tone 2) produced by a lower-frequency voice. In addition, contours are relatively flatter and less distinguishable in fluent speech. Production studies demonstrate that Mandarin tones vary according to the adjacent tones in running speech, and the amount of deviation depends on the nature of the tonal context (Xu, 1994, 1997). The perceptual challenge for listeners is to uncover the intended lexical tone in the face of this acoustic variability. Mirroring the literature for phonetic categorization, several studies have investigated how the perceptual system might "normalize" lexical tones across voices.

In an early study exploring tone normalization, Leather (1983) tested perception of syllables produced with Mandarin's tone 1 and tone 2 following natural sentences spoken by two different speakers. Native Mandarin Chinese listeners labeled acoustically identical target syllables with different lexical tones when targets were preceded by sentence contexts from different speakers, leading Leather (1983) to suggest that perception of lexical tones depends on perceived speaker identity. However, the descriptive data analysis showed that the influence of context was inconsistent across listeners and no inferential statistical analysis was applied because of the small ($N=5$) sample. In addition, little information of the direction of perceptual shift was provided, making it difficult to judge the directionality of any potential influence of context on lexical tone perception from these data.

More recent studies have examined the effect of context on Mandarin tone perception in paired syllables (Lin and Wang, 1985; Fox and Qi, 1990). In Lin and Wang's (1985) study, the *f*0 of the first syllable was held constant with a typical tone 1 *f*0 value (high-level) while the onset *f*0 of second syllable was manipulated across frequencies. Native Mandarin Chinese participants identified the tone of first syllable as tone 1 (high-level) or tone 2 (mid-rising). Results showed that as the onset *f*0 of the second syllable increased, participants were more likely to label the first syllable as tone 2. Using a similar paradigm, Fox and Qi (1990) examined perception of a series of Mandarin Chinese syllables that varied in *f*0 onset frequency from tone 1 to tone 2 in isolation and paired with a preceding syllable with a fixed *f*0 typical of tone 1 or tone 2. Native Chinese and native English listeners identified the tone category of second syllable.[1] Contrasting with Lin and Wang's (1985) findings, there was only a small and inconsistent difference between perception of syllables in isolation and in context and the effect was assimilatory, not contrastive.[2] Interestingly, the observed perceptual pattern was similar for both native (Mandarin) and non-native (English) participant groups.

These early studies examined perception of contour tones, which varied in both *f*0 frequency (height) and *f*0 contour (the change of *f*0) across the syllable. Since the results provided fairly inconsistent evidence for context-dependence, some researchers have suggested that listeners may rely mostly on intrinsic *f*0 contour information of the target syllable and much less on extrinsic information from context in perceiving contour lexical tones (Moore and Jongman, 1997).

Effects of context are much more evident among level tones that vary in *f*0 height but have similar contours. Two studies (Wong and Diehl, 2003; Francis *et al.*, 2006) have provided clear demonstrations of context-dependence in Cantonese level lexical tone perception. Wong and Diehl (2003) examined three level tones from Cantonese (tone 1: high-level tone, tone 3: mid-level tone, and tone 6: low-level tone; with relatively flat *f*0 contours and differentiated primarily by mean *f*0 frequency) as target stimuli. When listeners judged the identity of these tones in manipulated natural speech contexts, the same target stimuli were identified as tone 1 (high-level) 99.5% of the time with a low-frequency context and tone 6 (low-level) 95.8% of the time with a higher-frequency context. The same stimuli were identified as mid-level tone 3 91.9% of the time when the context had an intermediate mean *f*0 frequency. Francis *et al.* (2006) used a similar paradigm and also found that target stimuli were more likely to be perceived as a low level tone with a high-frequency synthesized context whereas the same stimuli were perceived as a high level tone with a lower-*f*0 synthesized context.

In Mandarin tones, Moore and Jongman (1997) examined perception of syllables varying from tone 2 (mid-rising) to tone 3 (low-falling-rising). When spoken in isolation, these tones have similar *f*0 contours, but differ in *f*0 height, the mean *f*0 frequency across the syllable. Preceding sentences recorded from two different speakers with different mean *f*0s, *f*0 turning points (the duration from syllable onset to the point of change in *f*0 direction) and Δf0s (the difference in *f*0 from onset to turning point) exerted a strong influence on Mandarin tone perception. Specifically, whereas perception of target stimuli varying only in *f*0 turning point was not influenced by preceding sentences, perception of tar-

gets varying in $\Delta f0$ was strongly context-dependent. Moreover, the effect was contrastive with respect to mean $f0$.[3] Stimuli were identified as tone 2 (mid-rising) in a low $f0$ speaker context and identification shifted to tone 3 (low-falling-rising) when there was a high $f0$ context. Since context sentences were recorded from two different talkers, Moore and Jongman (1997) argued that listeners use extrinsic $f0$ information from the context sentence to identify a speaker and this information exerts an influence on lexical tone identification. By this view, context-dependence in tone perception arises as a talker-contingent process, presumably mediated through a representation of speaker identity. Another possibility, motivated by analogy to the spectrally-contrastive shifts in phonetic categorization reviewed above and investigated in the present work, is that the different mean $f0$ frequencies of the speakers' voices can exert a contrastive influence on lexical tone perception independent of perception of speaker identity.

In summary, there are clear speaker normalization effects of level lexical tones for which $f0$ frequency is relatively constant little across the syllable (Wong and Diehl, 2003; Francis, *et al.*, 2006). The available evidence also suggests that the directionality of the influence of context on perception of level lexical tones is contrastive. As is observed for phonetic categorization in context, the spectra of preceding context affect perception of targets in a contrastive manner; higher-frequency contexts shift perception to lower frequencies whereas lower-frequency contexts cause the same targets to be perceived as a higher-frequency alternatives (Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al.*, 2006).

In comparison to the perceptual results with level lexical tones, observations of the influence of context on perception of contour lexical tones (distinguished by both $f0$ height and contour) have been much more mixed (Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990). Perception of contour tones could be much more dependent on intrinsic $f0$ information and much less affected by context than perception of level lexical tones. However, as noted above, there are some limitations in previous research investigating the effects of context on contour tones.

C. Research aims

The present research had several aims. Given the contradictory results in previous studies of lexical tone normalization for contour tones, we aimed to provide an unambiguous test of the influence of precursor context on contour tone perception. Here, we exploit paradigms similar to those used in studying effects of context on *level* tones (Wong and Diehl, 2003; Francis *et al.*, 2006) and stimuli similar to those of earlier studies of Mandarin *contour* tones (Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990). If perception of contour tones does rely on extrinsic information (i.e., it is sensitive to context), we expect to observe shifts in native Mandarin Chinese listeners' contour tone categorization as a function of preceding sentence contexts.

Another aim was to examine potential mechanisms of the pattern of perception that has been described as lexical tone normalization. Observing that natural precursor sentences recorded from different speakers shift tone perception, Moore and Jongman (1997) argued that the acoustic information of the context is used as a cue to identify the speaker and listeners' perception of lexical tones is calibrated via a representation of speaker identity. The pattern of perception thus is posited to be a result of talker-contingent processing. However, in light of the results reviewed above for context-dependent phonetic categorization, it is possible that these results arise not from perceived speaker identity but instead from spectral differences inherent in different speakers' utterances.

One possibility is that the effects may be a product of auditory, rather than phonetic or speaker-identity-specific, processing. Intriguingly, the context effects observed for lexical tones are contrastive in most studies: when there is a $f0$ context with a higher-frequency mean $f0$, the target is more likely to be labeled as a lower-frequency lexical tone and vice versa (Lin and Wang, 1985; Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al.* 2006). The contrastive directionality of these effects complements a wide range of studies of phonetic categorization whereby higher-frequency contexts shift perception toward a lower-frequency phonetic alternative (Ladefoged and Broadbent, 1957; Mann, 1980; Lotto *et al.*, 1997; Holt *et al.*, 2000; Holt and Lotto, 2002; Holt, 2005; see Lotto and Holt, 2006 for a brief review). In phonetic categorization, the effects of context are observed for single-syllable contexts (e.g., Mann, 1980; Lotto and Kluender, 1998; Holt *et al.*, 2000; Holt and Lotto, 2002) and also across sentence-length contexts (Ladefoged and Broadbent, 1957; Watkins and Makin, 1994, 1996; Holt, 2005, 2006a, 2006b) mirroring the temporal course of level tone context effects for syllables and sentences observed in the tone normalization literature (Lin and Wang, 1985; Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al.*, 2006). General auditory, rather than phonetic, mechanisms have been implicated in phonetic context effects because non-speech contexts mimicking the spectral properties of the speech contexts (but eliminating phonetic and articulatory information) produce similar context effects on speech targets (e.g., Lotto and Kluender, 1998; Holt, 1999, 2005, 2006a, 2006b; Fowler *et al.*, 2000; Holt *et al.*, 2000; Holt and Lotto, 2002; Coady *et al.*, 2003; Lotto *et al.*, 2003; Aravamudhan *et al.*, 2008). Further implicating general perceptual, rather than speech-specific mechanisms, such effects have been observed for a nonhuman animal species (Lotto *et al.*, 1997). The spectrally-contrastive directionality of context effects on lexical tone perception invites the possibility that general auditory processes may play a role in producing the patterns of perception that have been described as tone normalization.

Previous research does provide some clues to mechanism. Fox and Qi (1990) demonstrated that native Mandarin and non-native English listeners exhibit similar effect patterns for Mandarin tone perception. Wong (1998) also found that the effect of context could be observed with an English precursor (varying in $f0$ frequencies but not in lexical $f0$ information) for Cantonese and English bilinguals, although the effect was smaller than that elicited by the Cantonese

precursor contexts. Thus, it appears that linguistic experience with lexical tone may not be essential to these effects and that the influence of context does not rely on *f*0 conveying lexical information in context. Each of these findings is compatible with a spectral contrast account of lexical tone normalization, but they are not definitive. Francis *et al.* (2006) found that listeners' tonal judgments were proportional to the mean *f*0 frequency shifts of context sentences, consistent with predictions from spectral contrast. However, they also reported little effect of context on lexical tone perception using an unintelligible context precursor created by extracting the *f*0 contour of a sentence and using a "hummed" neutral vocal tract, a result that is unexpected from predictions of spectral contrast but that may be understood by more direct understanding of the spectral characteristics of precursor and target and their interactions.

In the present experiments, we sought to test directly whether patterns of perception considered to be examples of lexical tone normalization can be elicited with non-speech precursors. Drawing from the patterns of perception observed for context-dependent phonetic categorization (e.g., Holt, 2005), we predict that if general auditory processing plays a role in what has been described as speaker normalization of lexical tone, speech and nonspeech contexts that share energy in the region of *f*0, but eliminate speaker-specific and speech-specific information, should elicit similar context effects on Mandarin tone perception. Testing this hypothesis allows us to extend investigation of the role of spectral contrast in speech perception beyond English (e.g., Lotto and Kluender, 1998; Holt, 2005) to native Mandarin listeners and beyond segmental categorization to suprasegmental perception.

II. EXPERIMENT 1

The purpose of the first experiment is to extend the findings of speaker normalization of level tones (Wong and Diehl, 2003; Francis *et al.*, 2006) to perception of contour tones differing in both *f*0 height and contour. Mandarin's tone 1 (high-level) and tone 2 (mid-rising) differ in *f*0 height and contour and have been investigated in previous studies of tone normalization with mixed results (Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990). Whereas previous studies of tone normalization have used speech recorded from different speakers as contexts, the current study examines Mandarin listeners' perception of tone 1 and tone 2 in the context of a preceding Mandarin sentence from a single talker for which *f*0 has been manipulated. This approach to manipulating context holds all potential acoustic cues to speaker identity, other than mean *f*0, constant.

A. Method

1. Participants

Sixteen adult native Mandarin Chinese speakers participated in the experiment for a small payment. Participants did not learn any other Chinese dialects until 2 years old and had been in the United States for fewer than 5 years at the time the experiment was conducted. None reported any speech or hearing disability. Previous studies have shown that lexical tone processing is lateralized (Wang *et al.*, 2001; Wang *et al.*, 2004), so participants were given the Edinburgh handedness inventory before the experiments (Oldfield, 1971). Only right-handed listeners (inventory scores are no less than 40 out of 50) were included in the experiment to increase participant homogeneity.

2. Stimuli

The context stimuli were derived from a digital recording of a male native Mandarin speaker who spoke no other Chinese dialects (22050 Hz sampling rate, 16 bit resolution) uttering the Mandarin sentence: 请说这词/qing3 shuo1 zhe4 ci2 (please say this word). This semantically-neutral sentence was chosen because it possesses all four Mandarin tones.[4] Across five recorded utterances, the speaker's mean *f*0 was 159 Hz with an average range 117 Hz–211 Hz; a single sentence was chosen based on its clarity as judged by the first author, a native speaker of Mandarin Chinese. This sentence had a natural mean *f*0 of 162 Hz with a range 114–217 Hz. Two versions of the sentence were created by shifting the entire *f*0 contour of the sentence such that the average *f*0 of the high-frequency context stimuli was 200 Hz and the average *f*0 frequency of low-frequency context was 165 Hz (PRAAT 4.0, Boersma and Weenink, 2009). These two *f*0 frequencies were the onset *f*0 frequency values measured from recordings of the same speaker uttering tone 1 and tone 2 target stimuli (see below).

Three Mandarin syllables, /wu/, /yi/, /yü/, were used as target syllables. These syllables were chosen because they have different meanings when spoken in tone 1 and tone 2[5] (tone 1: room, cloth, dull; tone 2: none, wonder, fish). The target stimuli were derived from natural recordings of the same speaker who recorded the context sentences. One utterance per syllable was selected from five recorded tone 1 utterances, based on duration (around 450 ms) and clarity. An eight-step series varying perceptually from tone 1 to tone 2 was created for each syllable by manipulating the onset *f*0 frequency from 200 to 165 Hz in 5 Hz steps (PRAAT 4.0, Boersma and Weenink, 2009). From these onset values, *f*0 frequency transitioned linearly to an offset frequency of 200 Hz.[6]

Context sentences and target syllables were matched in rms amplitude. Each target stimulus was concatenated with each context sentence using MATLAB 7.0.1 (Matworks, Inc.), creating 48 stimuli, each with a total duration of 1750 ms. Figure 2(a) depicts the construction of stimuli and Fig. 2(b) gives a representative spectrogram for one stimulus (high frequency, 200 Hz mean *f*0, context and /wu/ target with lowest onset *f*0, 165 Hz).

3. Procedure

Each listener participated in two experiment sessions. In the experiment 1a, participants identified isolated target stimuli with no context. Each of the 24 (eight-step series $\times 3$ syllables) target stimuli was presented ten times in a random order. On each trial, a 500 ms fixation was shown on the screen before syllable presentation. Participants categorized the syllable by pressing "1" or "2" (tone 1 or tone 2) on

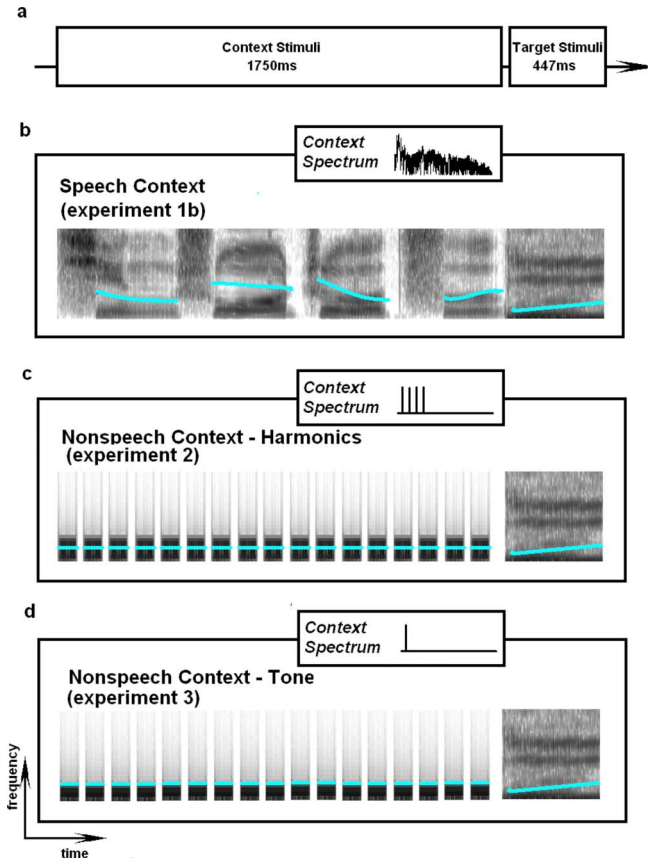

FIG. 2. (Color online) Schematic illustration of stimulus components (a) and representative spectrograms in time × frequency scales for high mean conditions of Experiment 1b (b), Experiment 2 (c), and Experiment 3 (d). The insets in panels (b)–(d) illustrate the spectra of the context sounds in frequency × amplitude axes. The spectrum of each experiment is showed in the inset in a frequency × amplitude scale.

the keyboard using the right hand. Experiment 1b followed Experiment 1a after a short break. The procedure was the same except syllable targets were preceded by context sentences varying in mean $f0$ frequency.

Acoustic presentation was under the control of E-prime (Schneider *et al.*, 2002); stimuli were presented diotically over linear headphones (Beyer DT-150) at approximately 70 dB SPL(A). Participants were tested in individual sound-attenuated booths. The experiment lasted approximately 1 h.

B. Results

Experiment 1a ensures that native Mandarin Chinese participants are able to categorize the syllable targets as Mandarin tone 1 and tone 2. Categorization was very regular across the three syllable series, so for this experiment and those that follow the data were collapsed across the /wu/, /yi/, /yü/ syllables. Average categorization responses as a function of target $f0$ onset frequency are shown in Fig. 3. A repeated measures analysis of variance (ANOVA) revealed a significant main effect for $f0$ onset frequency across the target-syllable series, $F(7,15)=326.74$, $p<0.01$, indicating that tone 1 and tone 2 were well categorized and demonstrating that manipulation of $f0$ onset frequency was sufficient to reliably shift Mandarin tone perception of isolated syllables.

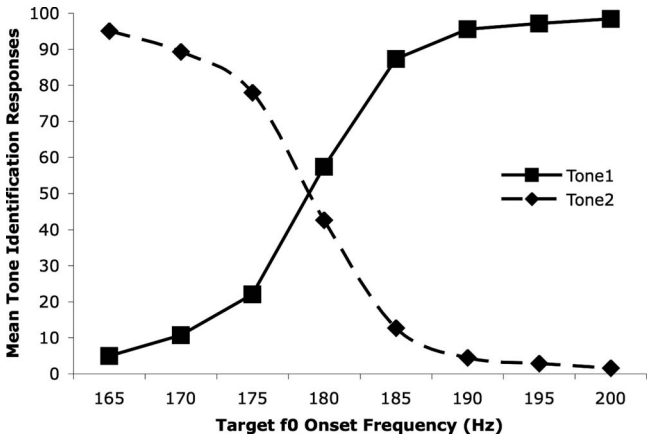

FIG. 3. Results of Experiment 1a. The tone stimuli were well-categorized across the tone 1 to tone 2 series.

Examination of individual's data ensured that each participant exhibited regular categorization across the $f0$ onset frequency series.

Results of Experiment 1b are presented in the Fig. 4(a) (top panel) as a function of target $f0$ onset frequency across participants. The solid line illustrates responses to the target in the context of higher-$f0$ precursor sentences whereas the dashed line represents the lower-$f0$ context. A 2 (speech context mean $f0$ frequency) × 8 (target $f0$ onset frequency) repeated measures ANOVA reveals a significant main effect of mean context $f0$, $F(1,15)=13.34$, $p<0.01$. There was also a main effect of target syllable $f0$ onset frequency, $F(7,15)=268.71$, $p<0.01$, as would be expected for orderly categorization across the tone series. The interaction between average context $f0$ frequency and target $f0$ onset frequency was significant, $F(7,15)=7.26$, $p<0.01$, indicating that the effect of context was exerted mainly in the middle of the tone series where the target stimuli were most perceptually ambiguous.

As expected from previous results and from parallels to context-dependence observed in phonetic context effects, the influence of context is contrastive: the high-frequency context (200 Hz mean $f0$) leads to more tone 2 responses (low onset $f0$), whereas the low-frequency context (165 Hz mean $f0$) leads to fewer tone 2 responses (more tone 1, high-frequency onset $f0$, responses). The significant influence of context indicates that Mandarin listeners do make use of context in contour tone perception. Contour tones, even though they can be distinguished by both $f0$ height and contour, are also context-dependent such that categorization is influenced by the mean $f0$ across a preceding context sentence. The context stimuli of the present experiment differ from those of many previous investigations of the influence of context on lexical tone (Leather, 1983; Moore and Jongman, 1997) because mean $f0$ frequency was manipulated independently of other speaker-specific acoustic variation. Mean $f0$ frequency therefore appears to drive the observed context effect observed here.

It is interesting to note that although the context sentences were manipulated so that they sampled different ranges of $f0$, few participants reported perceiving them as two different speakers. Yet, the contexts exerted an effect on

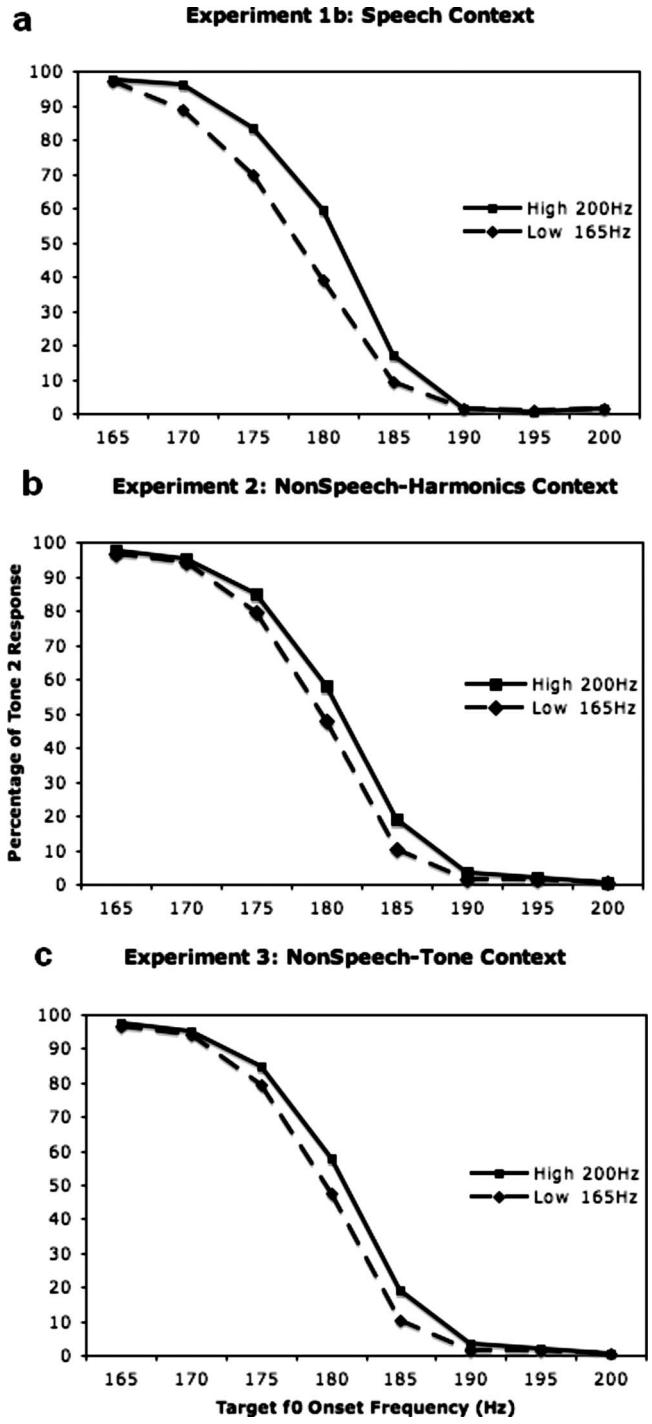

FIG. 4. Mean percentage of tone 2 responses for Experiment 1b (top panel), Experiment 2 (middle panel), and Experiment 3 (bottom panel).

tone perception. The data from 3 participants who reported perceiving different speakers were not qualitatively different from the remaining participants who heard the contexts as instances from the same speaker. These qualitative responses invite the question of whether it is necessary to perceive differences in speaker identity to normalize tone perception, as proposed by Moore and Jongman (1997). Moreover, the contrastive direction of the effect supports the possibility that patterns of perception described as speaker normalization for lexical tones may arise from general auditory mechanisms, as demonstrated for phonetic context effects (e.g., Lotto and Kluender, 1998; Holt, 2005).

III. EXPERIMENT 2

Experiment 2 tests this possibility by investigating Mandarin listeners' tone perception for syllables preceded by non-speech contexts that mimic some of the spectral characteristics of the sentence contexts used in Experiment 1. If general auditory processing accounts for context effects in lexical tones, non-speech contexts should elicit an influence on perception of tone in Mandarin syllables that parallels the sentence *f*0 contours they model. However, if speaker-contingent processing is important, there should be no observed effects of context because the non-speech context stimuli carry no information for speaker identity.

Previous studies (Holt, 2005, 2006a, 2006b) showed that the mean frequency of a sequence of non-speech acoustic stimuli exerts a spectrally-contrastive influence on subsequent speech categorization. In Experiment 1, manipulating the mean *f*0 of the precursor sentence was sufficient to shift listeners' tone 1 versus tone 2 perception, supporting the findings of previous studies of context-dependent lexical tone perception (Francis *et al.*, 2006) and verifying that such context-dependence exists for contour lexical tone. In another study suggestive of an influential role for mean *f*0 in tone perception, Francis *et al.* (2006) examined Cantonese tone perception in the context of monotone speech synthesized by setting *f*0 to a constant frequency, creating a "robot-like" timbre. The monotone contexts differing in mean *f*0 exerted a significant influence on Cantonese tone perception, strongly suggesting that listeners do not require the whole range of variation of speakers' *f*0, but rather rely on average *f*0. Experiment 2 further investigates the role of mean *f*0 in tone perception by utilizing complex *non-speech* stimuli composed of four sine-wave harmonics sharing the same *f*0 as the sentence contexts of Experiment 1b. Using these complex non-speech stimuli that do not possess speaker-identity information or information from which to recover speech gestures provides the opportunity to investigate the possibility that general auditory processing plays a role in tone normalization.

A. Method

1. Participants

The same group of 16 native Mandarin speakers from Experiment 1 was recruited to participate for a small payment. Participants returned for Experiment 2 at least 10 days after Experiment 1.

2. Stimuli

Two non-speech context stimuli were created using MATLAB 7.0.1 (Matworks, Inc.). The stimuli possessed the same mean *f*0 frequency as the sentence contexts in Experiment 1b. The high-frequency context had a 200 Hz *f*0 and the low-frequency context had a 165 Hz *f*0. Each non-speech context was a sequence of 17 tone complexes composed of four equal-amplitude sine-waves with frequencies at the first four multiples of the *f*0. Each of the 17 stimuli was 70 ms, with 5 ms linear amplitude ramps at onset and offset. The 17 non-speech tone complexes were separated by 30 ms silent intervals [see Fig. 2(c)].

These non-speech contexts preceded the target syllables from Experiment 1 with a 50 ms silence separating non-speech contexts and speech targets. The overall duration of the non-speech contexts was 1750 ms, the same duration as the speech contexts of Experiment 1b. Nonspeech context stimuli were matched in rms amplitude to the same value as speech context stimuli and target stimuli of Experiment 1.

3. Procedure

Experimental procedures were identical to those of Experiment 1b.

B. Results

Results are presented in the Fig. 4(b). A 2 (non-speech context $f0$) $\times$ 8 (speech target $f0$ onset frequency) repeated measures ANOVA analysis revealed a significant main effect of context $f0$: $F(1,15)=10.86$, $p<0.01$. As in Experiment 1b, there was a significant main effect of target $f0$ onset frequency, $F(7,15)=186.41$, $p<0.01$, indicating orderly categorization across the tone 1 to tone 2 target syllable series. There was also a significant interaction between context $f0$ frequency and target $f0$ onset frequency, $F(7,15)=2.49$, $p<0.05$, as expected for effects of context that exert the greatest influence on more perceptually-ambiguous targets. Again, the effect of context was contrastive: the non-speech context with a high $f0$ shifted listeners' lexical tone responses to tone 2 (lower-frequency $f0$ onset) whereas there were more tone 1 (higher-frequency $f0$ onset) responses when the non-speech context had a lower-frequency fundamental. Thus, stimuli that eliminate all potential speaker-specific and speech-specific information, but preserve spectral energy in the region of $f0$, are sufficient to shift Mandarin listeners' tone categorization when $f0$ frequency changes in the context. This finding suggests that general auditory processes may play a role in explaining what has been considered to be "speaker normalization" in lexical tones. Information about linguistic structure, speaker identity, or articulatory gestures does not appear to be necessary to account for the speaker normalization effect of sentence context observed in Experiment 1b.

IV. EXPERIMENT 3

The auditory system can extract pitch via the frequency of the fundamental and also via the intervals between higher-frequency harmonics of the fundamental (Bendor and Wang, 2005); the latter may be a more important manner of determining pitch in speech (Plack, 2005). The non-speech contexts in Experiment 2 possessed both types of information. Experiment 3 decouples these sources of pitch information to examine the influence of the first harmonic alone. Thus, this experiment is also an attempt to replicate the influence of non-speech context observed in Experiment 2 using an even simpler non-speech analog that is acoustically even less similar to the speech sentence contexts of Experiment 1b, yet preserves spectral energy in the region of $f0$. If such stimuli elicit a context effect on Mandarin listeners' tone perception, it provides stronger evidence for a role for general auditory processes in patterns of perception attributed to speaker normalization in lexical tone perception.

A. Method

1. Participants

The same group of sixteen native Mandarin speakers from Experiment 1 was recruited to participate for a small payment. They completed Experiment 3 after taking a short break following Experiment 2.

2. Stimuli

Two non-speech context stimuli were created using MATLAB 7.0.1 (Mathworks, Inc.). The stimuli had the same $f0$ frequency as the speech contexts of Experiment 1b and the non-speech four-harmonic stimuli of Experiment 2: the high-frequency context was composed of sine-waves of 200 Hz and the low-frequency context was made up of 165 Hz sine-waves. There were no high-frequency harmonics in the contexts of Experiment 3, leaving only a single sine-wave at the frequency of the first harmonic ($f0$), in contrast to the four-harmonic tone complexes of Experiment 2. Each context stimulus was composed of a sequence of 17, 70 ms sine-wave tones, each with the same frequency, with 30 ms silent intervals separating them [see Fig. 2(d)]. Each of the tones had linear 5 ms amplitude ramps at onset and offset.

These non-speech contexts preceded the speech syllable targets from Experiment 1 with 50 ms of silence separating speech and non-speech. As in the previous experiments, overall context stimulus duration was 1750 ms. Non-speech context stimuli were matched in rms amplitude to the value of speech context stimuli and target stimuli used in Experiment 1.

3. Procedure

Experimental procedures were identical to those of Experiment 1b.

B. Results

Data were scored as a function of target $f0$ onset frequency, collapsed across syllables. Results are plotted in the Fig. 4(c). A 2 (non-speech context $f0$ frequency) $\times$ 8 (speech target $f0$ onset frequency) repeated measures ANOVA revealed a significant main effect of non-speech context frequency on speech target categorization, $F(1,15)=14.21$, $p<0.01$. As expected, there was also a significant main effect of target $f0$ onset frequency, $F(7,15)=333.89$, $p<0.01$, indicating orderly categorization of targets and a significant interaction between context $f0$ frequency and target $f0$ onset frequency, $F(7,15)=4.55$, $p<0.01$, consistent with context exerting the greatest influence on perceptually-ambiguous mid-series targets. Once again, the observed effect was contrastive: the non-speech context with a higher-frequency $f0$ predicted more tone 2 (low $f0$ onset frequency) responses whereas the non-speech contexts with a lower-frequency $f0$ predicted more tone 1 (high $f0$ onset frequency) responses. Providing further evidence for a general auditory account of

what has been considered speaker normalization for tone perception, single sine-waves repeated across a sentence-length duration were sufficient to shift Mandarin listeners' tone perception. The directionality of the effect of the non-speech contexts mirrors the influence of the sentences they model in mean *f*0 frequency. Moreover, it appears that pitch information conveyed by the frequency interval between harmonics, present in the context stimuli of experiment 2 and absent in experiment 3 contexts, is not necessary to elicit an effect of non-speech context on speech. Thus, the spectral energy in the region of the mean fundamental frequency appears to be a key characteristic predicting the influence of context on identification of lexical tones.

V. GENERAL DISCUSSION

There were two main purposes for the present work. First, this study sought to determine whether there is evidence for context-dependent perception of contour tones. Previous research had provided strong evidence for context-dependence with level lexical tones that possess very similar *f*0 contours (Wong and Diehl, 2003; Francis *et al.*, 2006), but there were mixed results for tones differing along both *f*0 height and contour dimensions (Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990). Moore and Jongman (1997) suggested in their introduction that contour tones are perceived more according to their intrinsic *f*0 characteristics than level tones and are therefore being less susceptible to the influence of preceding speech context. Experiment 1b provides clear evidence that contour tones, in fact, are susceptible to the influence of context. Mandarin listeners do appear to use context in shaping their perception of contour tones, even when the tones may be distinguished by both *f*0 height and contour dimensions. Moreover, since the context sentences of Experiment 1b were created by manipulating the mean *f*0 of a sentence spoken by a single talker, all acoustic characteristics to speaker identity were held constant except for mean *f*0. This provides support for the possibility raised by Francis *et al.* (2006) that the average *f*0 of speech contributes to effects on tone perception more than does the range of variation of *f*0.

The current findings support Leather's (1983) argument that contour tones are context dependent. However, the context effect observed for Mandarin contour tones in Experiment 1b is much smaller than those observed for Cantonese level tones in previous studies. Identification of level tones can be almost completely (nearly 100%) shifted between two tone labels as a function of context (Wong and Diehl, 2003; Francis *et al.*, 2006). Previously-observed context-dependent shifts of Mandarin contour tone identification were also very reliable, but more modest than those found for Cantonese level tones. Moore and Jongman (1997) observed the largest average perceptual shift around 40% across contexts. The current experiment exposed an identification shift of around 20% for the most ambiguous target stimulus (*f*0=180 Hz) across contexts. It is interesting to note that stimuli in the Cantonese studies (Wong and Diehl, 2003; Francis *et al.*, 2006) had nearly identical tone contours whereas Moore and Jongman (1997) used similar, but not completely identical, tone contours. The *f*0 contours of stimuli in the current experiment are very different across targets. Thus, it seems that effects of context may be greater across target stimuli with more similar *f*0 contours. Although the current data provide strong support for context-dependence in perception of contour tones, the results should be interpreted in light of this pattern of observations. Perception of lexical tones, which may be distinguished by both *f*0 height and contour, appears to make use of both intrinsic and extrinsic context information and the degree of context-dependence may rely on the similarity among existing tones in the language. The reason for the observed pattern might come from the multi-dimensional nature of contour tones (e.g., Chandrasekaran *et al.* 2007; Barrie, 2007). If target tones can be well distinguished by *f*0 contours, context may be less necessary in establishing the percept. In other words, although contour tones exhibit significant speaker normalization or context-dependence, perception of them may be less susceptible to effects of context than level tones. On the other hand, it is possible that in the current study, listeners relied on the static cue (i.e., onset *f*0) more than the dynamic cue (i.e., contour) to make the tone decisions. If this were the case, the relatively smaller context-dependent perceptual effects for contour lexical tones compared to level tones (e.g., in Cantonese) may have arisen because level tones are better-differentiated by static acoustic cues like onset or offset *f*0 frequency than are contour tones. It is still unclear whether perception of onset *f*0 or *f*0 contour was influenced by context in current study. Further studies will be needed to address this question.

The second purpose of the present work was to examine the underlying mechanisms of patterns of perception described as speaker normalization effects in lexical tones. Most previous studies (Lin and Wang, 1985; Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al.*, 2006) observed a contrastive effect of speech context on lexical tone perception. With high-frequency *f*0 speech contexts, target stimuli were more likely to be perceived as a low-frequency tone and vice versa. The data from Experiment 1b replicated this finding for the influence of speech contexts on Mandarin contour tone perception. Experiments 2 and 3 found that two different types of non-speech contexts with spectral energy in the region of *f*0 exert contrastive context effects mirroring those observed for speech contexts in Experiment 1b after which they were modeled. Even single sine waves with frequencies at the mean *f*0 of the sentence contexts of Experiment 1b were sufficient to elicit an effect of context on lexical contour tone perception. These results suggest that linguistic information is not necessary in producing the kinds of context-dependent shifts in lexical tone categorization that have been called speaker normalization. This is consistent with the previous finding that an English precursor sentence elicited a context effect in labeling Cantonese lexical tones for Cantonese and English bilinguals (Wong, 1998). Contrary to speaker-contingent accounts (Moore and Jongman, 1997), it does not appear necessary to preserve information about speaker identity to observe these effects. Moreover, gestural information about vocal tract source appears not to be necessary to context-dependent patterns of percep-

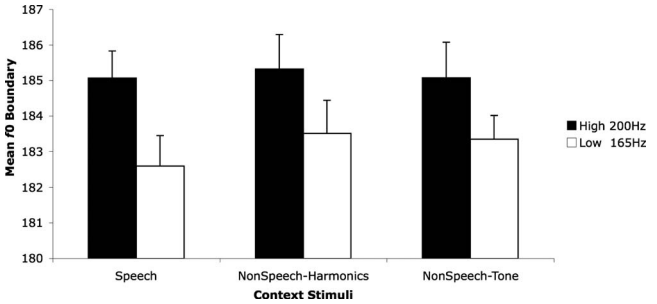

FIG. 5. Probit results for three context experiments (Experiment 1b, Experiment 2, and Experiment 3). A higher frequency $f0$ boundary indicates more tone 2 responses. Error bars indicate standard error of the mean.

tion thought be instances of normalization for lexical tone (see Lotto and Holt, 2006). Given that the only similarity across the contexts of Experiments 1b, 2, and 3 was the spectral energy in the region of $f0$, general auditory mechanisms are implicated.

To give a clear overview of the data across the experiments, Fig. 5 summarizes the results using probit analysis, a model of estimation for discrete decision variables (Finney, 1971), to estimate the identification boundaries and their shifts as a function of preceding context. To calculate the probit boundaries, a cumulative normal curve was used by transforming the percentage of "tone 2" responses to z-scores and finding the best fitting line via linear regression. The boundary was taken to be the $f0$ onset frequency of the target syllable corresponding to 50% on this line.

Since the three experiments share the same group of participants, a 3 (context stimulus type: speech, non-speech harmonics, non-speech single tones) $\times 2$ (context frequency: high, low) repeated measures ANOVA was conducted on the probit boundary values. The analysis confirms a significant context effect, $F(2,15)=46.242$, $p<0.001$ of the high vs low $f0$ contexts, with no significant main effect across the three types of context stimuli, $F(2,15)=0.651$, $p=0.53$. Each of the three context types produced a contrastive context effect whereby the high-frequency context led to a higher $f0$ boundary (i.e., a greater proportion of tone 2, low-frequency onset $f0$, responses) and vice versa. Of note, there was no significant interaction, $F(2,15)=1.305$, $p=0.35$, indicating that the influence of context on lexical tone identification was statistically indistinguishable in magnitude across the speech, non-speech harmonic; and non-speech tone context types. In sum, the probit analysis supports the prospect that the context-dependent shifts in lexical tone identification that have been attributed to speaker normalization may have their bases in general auditory mechanisms that produce spectral contrast.

Moore and Jongman (1997) suggested that because the contexts in their studies were recordings of natural speech from two speakers, tone perception might be mediated through a representation of speaker identity. However, there was no explicit test to verify that listeners indeed perceived the sentences as originating from two different talkers and there were no explicit tests of whether perceived speaker identity was key to eliciting the pattern of perception described as speaker normalization. Given the absence of information about speaker identity in the non-speech contexts of Experiments 2 and 3, it appears unlikely that speaker identity is a necessary factor in producing the kinds of categorization shifts that have been attributed to speaker normalization. Interestingly, the results of Francis *et al.* (2006) suggest that continuity of speaker identity is not necessary for tone normalization; evidence of context-dependent tone perception was even stronger when target and context stimuli came from different speakers, compared to when they were matched.

Other work suggests a role for influences of speaker identity in perception. Magnuson and Nusbaum (2007), for example, found that there were performance costs of adjusting to speaker variability when participants expect multiple speakers, whereas participants did not show this kind of performance when they heard the same materials but were expecting a single speaker. However, the present results demonstrate that auditory interactions of the spectra of context and target are sufficient to produce the kinds of identification shifts that have been taken as evidence of speaker-identity-driven mechanisms in previous research (e.g., Moore and Jongman, 1997).

If general auditory processes are primarily culpable, then one would expect commonality across languages. However, Jongman and Moore (2000) reported different patterns of normalization for Mandarin tone 2 and tone 3 for Mandarin and English listeners. Whereas Mandarin listeners' perception was influenced by preceding sentence contexts when context and target stimuli varied in the same acoustic dimension ($\Delta f0$ or $f0$ contour turning point), context only influenced English listeners when target stimuli varied in both $\Delta f0$ and $f0$ contour turning points. Jongman and Moore (2000) argued that language background aided Mandarin listeners in disambiguating phonemic contrasts, but normalization was the consequence of acoustic discriminability for English listeners. These results would seem to run counter to the present findings.

A major difference between Jongman and Moore (2000) and other cross-language studies in tone normalization (Fox and Qi, 1990; Wong, 1998) is that the former used lexical tones varying in both spectral ($\Delta f0$) and temporal ($f0$ contour turning point) dimensions. The current studies provide a strong support that tone normalization in spectral dimension may have its basis in general auditory processing. Of note, Wade and Holt (2005) showed that rate normalization effects can also be driven by sequences of sine-wave tones varying in their temporal characteristics. A possible explanation for the discrepancy between Jongman and Moore (2000) and other studies may be the interaction of spectral and temporal cues for lexical tone. English participants in Jongman and Moore's (2000) received short training in categorizing Mandarin tone 2 and tone 3. With covariance of $\Delta f0$ and $f0$ contour turning points as they learned these non-native categories, it is possible that English listeners could not separate these two dimensions. On the other hand, with much richer Mandarin lexical tone experience including using all four Mandarin tones, native speakers may be better able to use the cues independently. In other words, the influence of context that has been described as normalization may be driven by

common processes, but operative on very lexical tone categories with very different properties. This is an intriguing prospect that might be further tested.

The direction of the current work was motivated, in part, by previous studies of the influence of context on phonetic categorization. Those studies have produced three working conclusions: (1) Context-dependent phonetic categorization is contrastive in nature: higher frequency contexts shift perception toward lower-frequency targets and vice versa (Mann, 1980; Lotto *et al.*, 1997; Holt and Lotto, 2002; Holt, 2005; 2006a, 2006b; see Wade and Holt, 2005; Diehl and Walsh, 1989 for examples of temporally contrastive context-dependent phonetic categorization). (2) These effects can be elicited with non-speech stimuli modeling spectral/temporal characteristics of the speech contexts; thus, general contrastive perceptual mechanisms, rather than phonetic modules, gestural processing, or speaker-identity-driven mechanisms are implicated (e.g., Lotto and Kluender, 1998; Lotto and Holt, 2006). (3) In sentence-length acoustic materials for which the distribution of spectral energy varies, the key acoustic feature of context that influences the perception of a target sound is the mean frequency (Holt, 2005, 2006b).

The results of the current studies are consistent with each of these conclusions. As such, the present results extend findings of the influence of spectral contrast on speech perception cross-linguistically; previous studies have examined English whereas the present study investigates native Mandarin perception. Moreover, the present results broaden findings of spectral contrast to include suprasegmental, lexical tone. Most generally, the current studies suggest that contour lexical tones are not independent of extrinsic context and non-speech contexts influence lexical contour tone perception in a manner that mirrors the speech contexts they model with their spectra. Patterns of lexical tone perception that have been considered to be instances of speaker normalization may be driven, at least in part, by general auditory mechanisms that serve to perceptually exaggerate acoustic change, rather than speaker-identity-driven or articulatory-based processing.

ACKNOWLEDGMENTS

This work was supported by Grant No. R01DC004674 from the National Institutes of Health. J.H. received support from the Center for the Neural Basis of Cognition.

[1]English listeners were instructed to describe Tone 1 as high, unchanged (i.e., level) pitch and Tone 2 as a mid-rising pitch.

[2]The discrepant results between Lin and Wang (1985) and Fox and Qi (1990) may arise from methodology. See Moore and Jongman (1997) for a possible explanation.

[3]Moore and Jongman (1997) did not distinguish $\Delta f0$ and mean $f0$ in their studies. Context sentences from two speakers differed in both $\Delta f0$ and mean $f0$. Target stimuli had fixed onset and offset $f0$, but differed in both $\Delta f0$ and mean $f0$ because it is impossible to control mean $f0$ without varying $\Delta f0$. Other researchers have argued that the covariance of $\Delta f0$ and mean $f0$ is a weakness of this study (Wong and Diehl, 2003). However, although the influences of $\Delta f0$ and mean $f0$ cannot be decoupled in this study, it is the case that together they produced a significant, contrastive effect of context on lexical tone perception.

[4]Since it is impossible to rule out all contour information in the context, the current study used the context sentence 请说这词 which consisted of all four tones in Mandarin.

[5]Level tones are distinguished from contour tones in Cantonese because they have similar "level contour" and only the overall frequency distinguishes them. However, tone 1 is the only level tone in Mandarin so listeners can use both the level contour and overall frequency to distinguish tone 1 from other tones. In this case, "level" is considered as one of the types of $f0$ contours in Mandarin.

[6]Note that since the offset $f0$ frequency remains constant across target stimuli, changes in onset $f0$ frequency influence the slope of $f0$ frequency change across time.

Aravamudhan, R., Lotto, A. J., and Hawks, J. (**2008**). "Perceptual context effects of speech & non-speech sounds: The role of auditory categories," J. Acoust. Soc. Am. **124**, 1695–1703.

Barrie, M. (**2007**). "Contour tones and contrast in Chinese languages," J. East Asian Linguist. **16**(4), 337–362.

Bendor, D., and Wang, X. (**2005**). "The Neuronal representation of pitch in primate auditory cortex," Nature (London) **436**, 1161–1165.

Boersma, P, and Weenink, D.,

(**2009**). "Praat: doing phonetics by computer," (Version 4.0). http://www.praat.org/, (last accessed May 8, 2009)..

Chandrasekaran, B., Krishnan, A., and Gandour, J. (**2007**). "Mismatch negativity to pitch contours is influenced by language experience," Brain Res. **1128**, 148–156.

Coady, J. A., Kluender, K. R., and Rhode, W. S. (**2003**). "Effects of contrast between onsets of speech and other complex spectra," J. Acoust. Soc. Am. **114**, 2225–2235.

Diehl, R. L., and Walsh, M. A. (**1989**). "An auditory basis for the stimulus-length effect in the perception of stops and glides," J. Acoust. Soc. Am. **85**, 2154–2164.

Finney, D. J. (**1971**). *Probit Analysis* (Cambridge University Press, Cambridge, MA).

Fowler, C. A., Brown, J. M., and Mann, V. A. (**2000**). "Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans," J. Exp. Psychol. Hum. Percept. Perform. **26**, 877–888.

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

Francis, A., Ciocca, V., Wong, N., Leung, W., and Chu, P. (**2006**). "Extrinsic context affects perceptual normalization of lexical tone," J. Acoust. Soc. Am. **119**, 1712–1726.

Holt, L. L. (**1999**). "Auditory constraints on speech perception: An examination of spectral contrast," Ph.D. thesis, University of Wisconsin at Madison, Madison, WI.

Holt, L. L. (**2005**). "Temporally nonadjacent nonlinguistic sounds affect speech categorization," Psychol. Sci. **16**, 305–312.

Holt, L. L. (**2006a**). "Speech categorization in context: Joint effects of nonspeech and speech precursors," J. Acoust. Soc. Am. **119**, 4016–4026.

Holt, L. L. (**2006b**). "The mean matters: Effects of statistically-defined nonspeech spectral distributions," J. Acoust. Soc. Am. **120**, 2801–2817.

Holt, L. L., and Lotto, A. J. (**2002**). "Behavioral examinations of the level of auditory processing of speech context effects," Hear. Res. **167**, 156–169.

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. J., and Mullennix, J. W. (**1997**). *Talker Variability in Speech Processing* (Academic, San Diego).

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

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Ladefoged, P., and Maddieson, I. (**1996**). *Sounds of the World's Languages* (Blackwell, Oxford).

Leather, J. (**1983**). "Speaker normalization in perception of lexical tone," J. Phonetics **11**, 373–382.

Lin, T., and Wang, W. (**1985**). "Tone perception," J. Chin. Linguist. **2**, 59–69.

Lotto, A. J., and Holt, L. L. (**2006**). "Putting phonetic context effects into context: A commentary on Fowler (2006)," Percept. Psychophys. **68**, 178–183.

Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**, 602–619.

Lotto, A. J., Kluender, K. R., and Holt, L. L. (**1997**). "Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica)," J. Acoust. Soc. Am. **102**, 1134–1140.
Lotto, A. J., Sullivan, S. C., and Holt, L. L. (**2003**). "Central locus for non-speech effects on phonetic identification," J. Acoust. Soc. Am. **113**, 53–56.
Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. Hum. Percept. Perform. **33**, 391–409.
Mann, V. A. (**1980**). "Influence of preceding liquid on stop-consonant perception," Percept. Psychophys. **28**, 407–412.
Moore, C., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.
Oldfield, R. C. (**1971**). "The assessment and analysis of handedness: The Edinburgh inventory," Neurophysiology **9**, 97–113.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in the study of vowels," J. Acoust. Soc. Am. **24**, 175–184.
Plack, C. J. (**2005**). *The Sense of Hearing* (Lawrence Erlbaum Associates, Inc., London).
Schneider, W., Eschman, A., and Zuccolotto, A. (**2002**). "E-Prime user's guide," Psychology Software Tools Inc., Pittsburgh.
Wade, T., and Holt, L. L. (**2005**). "Perceptual effects of preceding non-speech rate on temporal properties of speech categories," Percept. Psychophys. **67**, 939–950.
Wang, Y., Jongman, A., and Sereno, J. (**2001**). "Dichotic perception of mandarin tones by Chinese and American listeners," Brain Lang **78**, 332–348.
Wang, Y., Behne, D., Jongman, A., and Sereno, J. (**2004**). "The role of linguistic experience in the hemispheric processing of lexical tone," Appl. Psycholinguist. **25**, 449–466.
Watkins, A. J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and for spectral-envelope distortion," J. Acoust. Soc. Am. **96**, 1263–1282.
Watkins, A. J., and Makin, S. J. (**1996**). "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**, 3749–3757.
Wong, P. C. M. (**1998**). "Speaker normalization in the perception of Cantonese level tones," MS thesis, University of Texas at Austin, Austin, TX.
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.
Xu, Y. (**1994**). "Production and perception of coarticulated tones," J. Acoust. Soc. Am. **95**, 2240–2253.
Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83.