LETTERS TO THE EDITOR

This Letters section is for publishing (a) brief acoustical research or applied acoustical reports, (b) comments on articles or letters previously published in this Journal, and (c) a reply by the article author to criticism by the Letter author in (b). Extensive reports should be submitted as articles, not in a letter series. Letters are peer-reviewed on the same basis as articles, but usually require less review time before acceptance. Letters cannot exceed four printed pages (approximately 3000–4000 words) including figures, tables, references, and a required abstract of about 100 words.

Evidence for the central origin of lexical tone normalization (L)

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 5 October 2010; accepted 29 December 2010)

Huang and Holt [(2009). J. Acoust. Soc. Am. **125**, 3983–3994] suggest that listeners may dynamically tune lexical tone perception via general auditory sensitivity to the mean *f*0 of the preceding context, effectively normalizing pitch differences across talkers. The present experiments further examine the effect using the missing-*f*0 phenomenon as a means of determining the level of auditory processing at which lexical tone normalization occurs. Speech contexts filtered to remove or mask low-frequency *f*0 energy elicited contrastive context effects. Central, rather than peripheral, auditory processes may be responsible for the context-dependence that has been considered to be lexical tone normalization. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3543994]

PACS number(s): 43.71.An, 43.66.Ba, 43.71.Hw [JES] Pages: 1145–1148

I. INTRODUCTION

Tone languages use variations in fundamental frequency (*f*0) to contrast meaning. For example, Mandarin Chinese has four different tones: high-level tone (tone 1), mid-rising tone (tone 2), low-falling-rising tone (tone 3), and high-falling tone (tone 4) that may shift the meaning of an utterance. However, what constitutes a particular tone class varies greatly since talkers differ considerably in their average *f*0. Thus, there is ambiguity for listeners of tonal languages: A particular *f*0 may signal a lower lexical tone for a relatively high-pitched female voice, whereas the same *f*0 may signal a higher lexical tone for a lower-pitched male voice.

A great deal of research has investigated how listeners "normalize" lexical tone perception to accommodate such talker differences (Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990; Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al*., 2006; Huang and Holt, 2009). In general, these studies have demonstrated that context-dependent perception helps to resolve the variability in the speech signal (e.g., Wong and Diehl, 2003; Francis *et al*., 2006; Huang and Holt, 2009). In a recent study, Huang and Holt (2009) found that native Mandarin listeners' perception of Mandarin first and second tones was affected by the mean fundamental frequency (*f*0) of a preceding sentence. The influence was contrastive: When target syllables were preceded by a context sentence with a higher mean f0, they were more often categorized as a lower *f*0 tone (mid-rising tone 2), whereas the same targets were more often identified as a higher *f*0 tone (high-level tone 1) when preceded by the same sentence with a lower mean *f*0.

Huang and Holt (2009) argued that the mechanisms underlying context-dependent lexical tone perception may have a general auditory, rather than speech-specific, basis. In support of this hypothesis, when Mandarin participants heard the target syllables preceded by sequences of sine-wave tones (or four-harmonic tone complexes) with *f*0 frequencies at the mean *f*0 of the high- and low-*f*0 context sentences, lexical tone categorization was context-dependent in the same manner as observed for sentence contexts. Targets were identified as the lower-*f*0 tone 2 more often following a higher-frequency sequence of sine-wave tones whereas they were more often categorized as higher-*f*0 tone 1 following low-frequency tones. In fact, the influence of these nonspeech contexts on lexical tone perception was statistically indistinguishable in magnitude from the influence of the naturally spoken sentence contexts. Since the nonspeech tone sequences possessed no articulatory or speaker-identity information, these results suggest a role for general auditory processing instead of speech- or speaker-specific mechanisms.

In the Huang and Holt (2009) experiments, the first harmonic (*f*0) was present for both speech and nonspeech contexts. However, *f*0 need not be present to elicit a strong pitch percept (Licklider, 1956; Plack *et al*., 2005; Yost, 2009). For example, a complex spectrum composed of 200, 300, and 400 Hz tones has a pitch of 100 Hz even without acoustic energy at 100 Hz. This *missing-f0* phenomenon is a central aspect of pitch perception and its existence provides the possibility of probing the basis of context-dependent tone normalization in speech processing. Since the lowest-frequency harmonics are absent in missing-*f*0 stimuli, pitch must be

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

based on the temporal regularities within a sound's waveform or the analysis of the match to spectral templates at a central level (see Yost, 2009 for review).

Here, we test whether missing-*f*0 contexts affect lexical tone categorization to investigate whether the auditory mechanisms involved in the context-dependent perception thought of as lexical tone normalization have a central origin. Studies of the missing-*f*0 have noted that *f*0 can be reintroduced as combination tone distortion products on the basilar membrane if they are simply filtered from a complex sound (Robles *et al*., 1991, 1997; Plack *et al*., 2005). In this case, the input to the central auditory system may possess a well-resolved *f*0 harmonic as a result of basilar membrane nonlinearities (although it was absent in the signal). To protect against this in the current experiments, two missing-*f*0 manipulations are used to eliminate *f*0 from speech contexts: high-pass filtering low-frequency energy (experiment 1b) and high-pass filtering combined with a low-frequency noise masker (experiment 1c).

II. EXPERIMENT

A. Method

1. Participants

Eleven adult native-Mandarin speakers were recruited for a small payment. Participants did not learn any other Chinese dialects until 2 yr old and had been in the United States for fewer than 5 yr at the time the experiment was conducted. None reported any speech or hearing disability and all were right-handed (Edinburgh handedness inventory no less than 40 out of 50; Oldfield, 1971).

2. Stimuli

The context stimuli were derived from a digital recording of the Mandarin sentence 请说这词/qing3 shuo1 zhe4 ci2 (*Please say this word*) recorded from a male native-Mandarin speaker who spoke no other Chinese dialects (22 050 Hz sampling rate, 16 bit resolution). This semantically neutral sentence possesses all four Mandarin tones and was used previously by Huang and Holt (2009). The sentence had a natural *f*0 mean of 162 Hz with a range of 114–217 Hz. From this recording, a high-frequency context stimulus with an average *f*0 of 200 Hz and a low-frequency context stimulus with an average *f*0 frequency of 165 Hz were created by shifting the entire *f*0 contour of the sentence (Praat Version 4.0, Boersma and Weenink, 2009). These two average *f*0 frequencies were selected based on the range of onset *f*0 frequencies of target stimuli (see below).

Two sets of missing-*f*0 context stimuli were created from these sentences. For experiment 1b, the context sentences were high-pass filtered at 300 Hz to remove the lowest-frequency harmonics of *f*0 [see Fig. 1(c); left panel]. For experiment 1c, a low-pass filtered white noise masker below 300 Hz was added to context stimuli in experiment 1b (speech high-pass filtered at 300 Hz). The noise masker was rms-matched in energy to the acoustic energy of speech sentences below 300 Hz so that the noise possessed acoustic energy equal to the low-frequency speech is removed with the high-pass filter [see Fig. 1(d); left panel].

Target stimuli were three Mandarin syllables, /wu/, /yi/, /yü/, as described by Huang and Holt (2009). Targets were derived from recordings of the talker who produced the context sentences and an eight-step series varying perceptually from Mandarin lexical 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 Version 4.0, Boersma and Weenink, 2009). The offset *f*0 frequency of each syllable was anchored at 200 Hz. Each context sentence and target syllable was matched in overall RMS amplitude, with 48 stimuli in both experiments 1b and 1c.

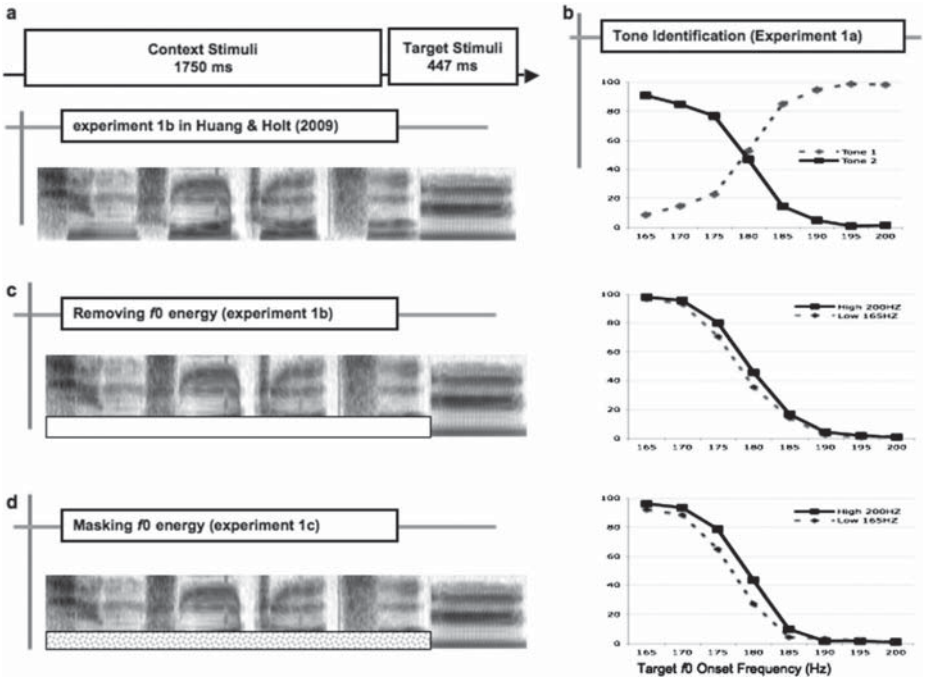

FIG. 1. (a) Schematic illustration of stimulus components and speech contexts of Huang and Holt's (2009) experiment 1b; (b) results of experiment 1a for tone 1 and tone 2 identification; (c) stylized illustration of stimulus manipulation and results of experiment 1b; left panel: spectrogram in time × frequency dimensions for the high *f*0 speech context; right panel: mean percentage of tone 2 responses; (d) stylized illustration of stimulus manipulation and results of experiment 1c; left panel: spectrogram in time × frequency dimensions for the high *f*0 speech context; right panel: mean percentage of tone 2 responses.

3. Procedure

A 500-ms fixation preceded auditory stimulus presentation. Native-Mandarin participants categorized the target syllable by pressing "1" or "2" (tone 1 or tone 2) on a keyboard using the right hand. Participants first identified isolated target stimuli in experiment 1a, responding to each target ten times in random order. After a short break, listeners identified the same syllable targets preceded by context sentences varying in mean *f*0 frequency (high/low) for high-pass filtered sentences with no acoustic information below 300 Hz (experiment 1b) and for the same high-pass filtered sentences with a low-pass white noise masker below 300 Hz (experiment 1c). Each context plus target stimulus was presented ten times in random order. The order of experiments 1b and 1c was counterbalanced across participants.

Acoustic presentation was under the control of E-prime (Schneider *et al.*, 2002); stimuli was presented diotically over linear headphones (Beyer DT-150, Beyer Dynamics, Farmingdale, NY) at approximately 70 dB sound pressure level (SPL) (A) over the course of a 1-h experiment that took place in individual sound-attenuated booths.

III. RESULTS

Figure 1(b) illustrates the average speech target categorization as a function of *f*0 onset-frequency in isolation, collapsed across the three syllable targets. A repeated-measures ANOVA (analysis of variance) revealed a significant main effect for *f*0 onset-frequency across the target-syllable series, $F\ (7,10) = 326.74$, $p < 0.01$, indicating good categorization and sufficient manipulation of *f*0 onset-frequency to shift Mandarin tone perception of /wu/, /yi/, /yü/ syllables from tone 1 to tone 2 (see also Huang and Holt, 2009). Individual data were conformed to this group pattern.

The right panels of Figs. 1(c) and 1(d) illustrate the influence of the mean *f*0 of speech context on target categorization when acoustic energy below 300 Hz was absent in experiment 1b and when acoustic energy below 300Hz was replaced by white noise in experiment 1c. A 2 (mean *f*0 frequency) × 8 (target *f*0 onset-frequency) repeated-measures ANOVA reveals a significant main effect of context *f*0 in both experiments: $F(1, 10) = 9.72$, $p < 0.05$ (experiment 1b) and $F(1, 10) = 15.50$, $p < 0.05$ (experiment 1c). Consistent with previous results for sentences possessing *f*0 (Huang and Holt, 2009), the influence of contexts was contrastive: Listeners reported more tone 2 responses (low-frequency onset *f*0) following the high-frequency context (200 Hz mean *f*0), and there were more tone 1 responses (high-frequency onset *f*0) in the low-frequency context condition (165 Hz mean *f*0). Tone perception shifted according to the preceding contexts even when low-frequency *f*0 energy below 300 Hz was completely removed or replaced by white noise.

There was also a main effect of target-syllable *f*0 onset-frequency, $F(7, 10) = 155.77$, $p < 0.001$ (experiment 1b) and $F(7, 10) = 192.69$, $p < 0.001$ (experiment 1c), confirming the orderly categorization across the tone series found in experiment 1a. The interaction between average context *f*0 frequency and target *f*0 onset-frequency was also significant, $F(7, 10) = 2.63$, $p < 0.001$ (experiment 1b) and $F(7, 10) = 2.63$, $p < 0.001$ (experiment 1c), indicating that context primarily influenced the most perceptually ambiguous syllables in the middle of target series.

The data of experiment 1b and 1c were merged for further analysis. A 2 (manipulation: removing low-frequency energy vs removing low-frequency energy and replacing with noise) × 2 (speech context mean *f*0 frequency) × 8 (target *f*0 onset-frequency) three-way repeated-measures ANOVA revealed a contrastive context effect of mean *f*0, $F(1, 10) = 18.55$, $p < 0.05$, as expected. In addition, there was no significant difference between removing *f*0 energy (experiment 1b) and replacing *f*0 energy by noise (experiment 1c), $F(1, 10) = 3.13$, $p = 0.11$; and the interaction between mean context *f*0 frequency and manipulations of the low frequency was not significant, $F(1, 10) = 1.66$, $p = 0.23$. The two versions of missing-*f*0 contexts did not differ in the extent to which they elicited context-dependent lexical tone categorization. Listeners appear to be able to use pitch information extracted at a central level to normalize lexical tone perception when the *f*0 is removed or even replaced by white noise.

Again, there was a main effect of target-syllable *f*0 onset-frequency, $F(7,10) = 217.62$, $p < 0.001$; and the interaction between the mean context *f*0 frequency and target *f*0 onset-frequency was also significant, $F(7, 10) = 6.93$, $p < 0.001$. The interaction of the missing-*f*0 manipulation and target *f*0 onset-frequency was not significant, $F(7, 10) = 0.93$, $p = 0.49$, indicating that target categorization was robust across different manipulations of removing low-frequency energy. The three-way interaction of context *f*0 manipulation, mean context *f*0 frequency, and target *f*0 onset-frequency was not significant, $F\ (7, 10) = 0.97$, $p = 0.46$.

IV. DISCUSSION

Mandarin lexical tone perception was significantly influenced even by contexts possessing no low-frequency energy in the range of *f*0. Moreover, the directionality of the effect was contrastive: There were more higher-frequency (tone 1) responses following missing-*f*0 contexts that could be resolved to possess lower-frequency pitch, whereas the same targets were more often perceived as lower-frequency (tone 2) when preceded by missing-*f*0 targets with higher pitch. This contrastive pattern mirrors results of talker normalization and tone normalization in previous studies (e.g., Ladefoged and Broadbent, 1957; Huang and Holt, 2009). In short, the contrastive tone normalization effect is robust even when *f*0 energy is completely removed and masked by noise.

The results of Huang and Holt (2009) suggest that a general auditory mechanism underlies context-dependent lexical tone perception thought to be indicative of talker normalization. The present results are consistent with a central auditory, rather than peripheral, origin for this mechanism because the two missing-*f*0 speech contexts (high-pass filtering low-frequency energy in experiment 1b and high-pass filtering combined with replacing low-frequencies with a noise masker in experiment 1c) similarly elicited contrastive effects on Mandarin tone targets. In addition, Huang and Holt (2009) suggests a shared mechanism underlying speaker normalization effects

for both suprasegmentals (i.e., lexical tone) and segmentals (i.e., vowels and consonants) because of the similar contrastive effect pattern in these studies (e.g., Moore and Jongman, 1997; Wong and Diehl, 2003; Francis *et al*., 2006; Huang and Holt, 2009; Ladefoged and Broadbent, 1957; Watkins and Makin, 1994, 1996).

Missing-fundamental pitch is also found among listeners without significant linguistic experience (animals: Tomlinson and Schwarz, 1988; human infants: Clarkson and Clifton, 1985). The present data predict on the basis of shared mechanisms of general auditory processing that animal and infant listeners also may exhibit lexical tone "normalization" for speech and nonspeech contexts. Rather than a speech-, talker- or articulatory-gestural-specific mechanism, lexical tone normalization may arise from the tendency for perceptual systems to emphasize change through perceptual contrast.

Boersma, P., and Weenink, D. (**2009**). "Praat: Doing phonetics by computer (Version 4.0) [Computer program]," http://www.praat.org (Last viewed September 12, 2010).

Clarkson, M. G., and Clifton, R. K. (**1985**). "Infant pitch perception: Evidence for responding to pitch categories and the missing fundamental," J. Acoust. Soc. Am. **77**, 1521–1528.

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**(3), 1712–1726.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**, 3983–3994.

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.

Licklider, J. C. R. (**1956**). "Auditory frequency analysis," in *Information Theory, Third London Symposium*, edited by C. Cherry (Butterworth Scientific, London), pp. 253–68.

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

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.

Robles, L., Ruggero, M., and Rich, N. (**1991**). "Two-tone distortion in the basilar membrane of the cochlea," Nature (London) **349**, 413–414.

Robles, L., Ruggero, M., and Rich, N. (**1997**). "Two-tone distortion in the basilar membrane of the chinchilla cochlea," J. Neurophysiol. **77**, 2385–2399.

Plack, C. J. (**2005**). *The Sense of Hearing* (Lawrence Erlbaum Associates, New York), pp. 132–151.

Schneider, W., Eschman, A., and Zuccolotto, A. (**2002**). "E-PRIME User's guide," (Psychology Software Tools Inc., Pittsburgh).

Tomlinson, R. W., and Schwarz, D. W. (**1988**). "Perception of the missing fundamental in nonhuman primates," J. Acoust. Soc. Am. **84**, 560–565.

Watkins, A. J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and for spectral-envelope distortion," J. Acoust. Soc. Am. **96**(3), 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**(6), 3749–3757.

Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intra-talker variation in Cantonese level tones," J. Speech Lang. Hear. Res. **46**, 413–421.

Yost, W. A. (**2009**). "Pitch perception," Attention Percept. Psychophys. **71**, 1701–1715.