The effect of context duration on Mandarin listeners' tone normalization

Xin Luo[a] **and Krista B. Ashmore**
Department of Speech, Language, and Hearing Sciences, Purdue University,
500 Oval Drive, West Lafayette, Indiana 47907
luo5@purdue.edu, kashmore@purdue.edu

Abstract: Tone normalization has been observed in Mandarin listeners, who contrastively adjust tone recognition using context pitch cues. This study tested the effect of context duration on Mandarin tone normalization. The target tones varied from Tone 1 (high-flat) to Tone 2 (mid-rising). The preceding phrase was modified to have different durations with 160- or 200-Hz mean fundamental frequencies ($F0$s). The results showed that the high-$F0$ context elicited significantly more Tone-2 responses than the low-$F0$ context, even when the contexts were 125 ms. The contrastive context effect saturated with the 250-ms contexts, indicating a 250-ms critical context duration for robust tone normalization.

PACS numbers: 43.71.An, 43.66.Ba, 43.66.Hg [DDO]
Date Received: April 28, 2014 **Date Accepted:** June 12, 2014

1. Introduction

The four lexical tones in Mandarin Chinese are used to contrast word meaning at a syllabic level (e.g., Chao, 1948). For example, the same Mandarin syllable /i/ with the four tones could mean cloth, aunt, chair, and meaning, respectively. The four Mandarin tones have the principal acoustic correlate of fundamental frequency ($F0$) contour (i.e., Tone 1: High-flat; Tone 2: Mid-rising; Tone 3: Low-falling-rising; Tone 4: High-falling), although vowel duration and amplitude envelope may also contribute to Mandarin tone recognition (e.g., Luo and Fu, 2004). The $F0$ cues to Mandarin tones vary greatly with different tonal contexts, individual speaker characteristics, and many other factors (e.g., Xu, 1997; Moore and Jongman, 1997). For example, a speaker with a high-pitched voice would produce a specific tone with higher $F0$s than a speaker with a low-pitched voice. To accurately identify the target tone in spite of the $F0$ variance, listeners have to perceive the pitch of the target tone relative to the voice pitch of the speaker (a process called speaker normalization for tone perception, or in short, tone normalization).

Behavior evidence of speaker normalization has been found for the recognition of Cantonese level tones (e.g., Francis *et al.*, 2006; Wong and Diehl, 2003) and Mandarin contour tones (e.g., Huang and Holt, 2009; Moore and Jongman, 1997). In these tone normalization studies, a preceding sentence with a high mean $F0$ (simulating a high-pitched voice) elicited more low tone responses for the same target tones than a preceding sentence with a low mean $F0$ (simulating a low-pitched voice). In other words, the $F0$ of the sentence context had a contrastive effect on tone recognition. Huang and Holt (2009) found that Mandarin tone recognition was also contrastively affected by the mean $F0$s of the non-speech contexts consisting of harmonic complex tones or pure tones. Thus, tone normalization does not need speaker or phonetic information and may arise from general auditory processing of pitch contrast between the

[a] Author to whom correspondence should be addressed.

context and target stimuli, as long as listeners have the linguistic experience of Mandarin tones (e.g., Luo and Ashmore, 2014).

Studies have also investigated the context pitch cues that were used by listeners to achieve tone normalization. For example, the contrastive effect of monotone contexts in Francis *et al.* (2006) suggests that exposure to the mean $F0$ rather than the full $F0$ range of a speaker was sufficient for Cantonese level tone normalization. Cantonese level tone normalization was also possible (although less optimal) when the $F0$ changes between the contexts were less than the typical $F0$ differences between the target level tones. In Wong and Diehl (2003), the preceding context was divided into two halves with separate $F0$ manipulations. It was found that Cantonese level tone normalization was mostly based on the $F0$ information in the second half rather than the first half of the context. Thus, tone normalization may use a moving window or running average mechanism for context pitch perception, which weighs recent pitch cues more heavily than earlier pitch cues.

In this study, we are interested in the duration of context needed for Mandarin tone normalization. As the context duration increases, more context pitch cues will be available for listeners to refer to during target tone recognition. It is possible that the context effect on tone recognition may saturate when the context is longer than a critical duration. In Luo and Ashmore (2014), context effects on Mandarin tone recognition were actually similar for both the 500-ms non-speech contexts and 1319-ms speech contexts with the same mean $F0$s. Eramela (2002) tested Cantonese level tone normalization with contexts of different durations, each containing a different number of syllables with the same mid-level tone. She found that the percentages of expected tone responses (e.g., low-level tone responses in a high-$F0$ context) significantly increased with context duration and reached a plateau with a 450-ms context (or with two preceding syllables). However, there are three caveats in her study that may influence the interpretation of the results. First, Eramela tested the different context durations in a fixed ascending order. The performance with longer contexts was always tested later and thus may have improved partially due to practice effects. Second, although the same level tone of the different preceding syllables helped maintain the pitch level with the different context durations, realistic contexts most often contain different tones with pitch variations. Third, because the number of syllables changed with the context duration, it is unclear whether it was the specific context duration (450 ms) or it was the specific number of syllables (two) that provided enough context pitch cues for tone normalization. In this study, the effect of context duration on Mandarin tone normalization was tested. Two preceding syllables had different Mandarin tones and their speaking rate was modified to create contexts of different durations (tested in random order). The context duration was thus varied without changing the number of syllables. The present results were compared to those of Eramela (2002) to shed light on the time course of tone normalization.

2. Methods

2.1 Subjects

Five female and 6 male native Mandarin normal-hearing adults ranging in age from 21 to 33 years with a mean age of 25 years were recruited from students at Purdue University. They started learning English in primary or middle school and have been in the United States for 2 to 5 years. For all subjects, hearing thresholds were below 25 dB hearing level at octaves between 0.25 and 8 kHz in both ears. This study was reviewed and approved by the Institutional Review Board of Purdue University. All subjects provided informed consent and were compensated for their participation.

2.2 Stimuli

All the speech stimuli had a 44 100-Hz sampling rate with 16-bit resolution and were manipulated using the PRAAT 5.3.17 software (Boersma and Weenink, 2012). The target Tone 1–Tone 2 (high-flat and mid-rising) series were the same as those in Luo and Ashmore (2014). A male native Mandarin speaker produced a Mandarin syllable /i/ in Tone 1. The 506-ms syllable was recorded and then manipulated to have 1 of 9 linear $F0$ contours with the onset $F0$ ranging from 160 to 200 Hz in 5-Hz steps and the offset $F0$ fixed at 200 Hz. These $F0$ values were modeled after the male speaker's natural productions of Tone 1 and Tone 2. A semantically neutral sentence "请听下个词" /qing3 ting1 xia4 ge4 ci2/, meaning "please listen to the next word," was also recorded from the male speaker. Instead of using the whole sentence as in Luo and Ashmore (2014), this study used the first two words of the sentence (i.e., "请听" /qing3 ting1/, meaning "please listen") to create the different preceding contexts. Compared to the 1319-ms sentence, the 517-ms two-word phrase was more suitable for creating the different preceding contexts, because its duration was in the middle of the five tested context durations (i.e., 125, 250, 500, 1000, and 1500 ms). The two words had two different Mandarin tones (the lowest Tone 3 and the highest Tone 1, respectively), which reduced the perceptual adaptation to any particular tone and well represented the pitch range of the male speaker (from 124 to 200 Hz with a mean of 169 Hz). The two-word phrase was first manipulated to have the five different durations without varying the formant frequencies and $F0$ contour. For each of the five durations, the entire $F0$ contour of the two-word phrase was then shifted up or down to create a high- or low-$F0$ context with a mean $F0$ of 200 or 160 Hz, respectively. These two mean $F0$s in the contexts corresponded to the highest and lowest target onset $F0$s. Thus, there were totally ten different contexts (five durations × two $F0$s). The context and target stimuli were matched in root-mean-square level and were concatenated with a 50-ms inter-stimulus interval to test tone recognition with context.

2.3 Procedures

The stimuli were presented via the basic loudspeaker of a GSI-61 audiometer at 70 dBA in a double-walled, sound-treated booth. Mandarin tone recognition was tested with a two-alternative, forced-choice task. Two buttons on a computer screen showed the two response choices for each trial. One button was labeled with "Tone 1" and a flat line, while the other with "Tone 2" and a rising diagonal line. Subjects chose the target tone by clicking on one of the two response buttons. The percentage of Tone-2 responses was recorded for each target stimulus in each context condition.

The task was explained and ten sample stimuli randomly selected from the test set were presented for practice before testing. In the first session, tone recognition was tested without context to make sure that the range of target onset $F0$ was sufficient for the identification of both Tone 1 and Tone 2. The 9 target stimuli were tested 10 times in random order, resulting in totally 90 trials in this session. The following five sessions tested tone recognition with contexts of the five different durations, respectively. The testing order of the five context durations was randomized across subjects to avoid any order effects. In each session with contexts of a particular duration, the 9 target stimuli preceded by either the high- or low-$F0$ context were tested 10 times in random order, resulting in totally 180 trials. The choice of the high- or low-$F0$ context was randomized from trial to trial. No feedback was provided during any session.

3. Results

Figure 1(a) shows the percentage of Tone-2 responses as a function of target onset $F0$ for tone recognition without context. All subjects had a typical S-shaped psychometric function and their responses gradually changed from Tone 2 to Tone 1 as the target onset $F0$ increased from 160 to 200 Hz. The response percentages were first transformed by a rationalized arcsine transformation (Studebaker, 1985) and then analyzed by a one-way repeated-measures (RM) analysis of variance (ANOVA) with target onset $F0$ as the factor. The percentage of Tone-2 responses without context was significantly affected by target onset $F0$ ($F_{8,80} = 141.65$, $p < 0.001$).

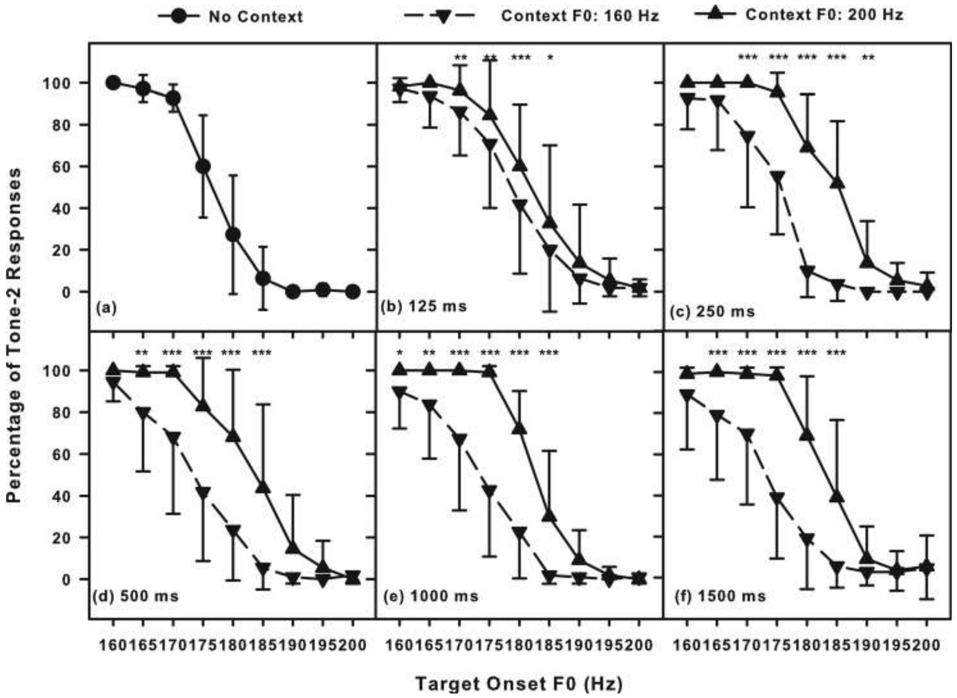

Fig. 1. Percentage of Tone-2 responses as a function of target onset $F0$ for tone recognition without context (a) and with the high- (upward triangles) or low-$F0$ contexts (downward triangles) of different durations from 125 to 1500 ms [(b)–(f), respectively]. Symbols represent the mean, while error bars represent the standard deviation across subjects. For clarity of illustration, error bars are shown in only one direction. Target tones with significantly different responses in the high- and low-$F0$ context conditions are indicated by asterisks in each panel (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

Figures 1(b)–1(f) show the percentage of Tone-2 responses as a function of target onset $F0$ for tone recognition with either the high- (upward triangles) or low-$F0$ contexts (downward triangles) of different durations from 125 to 1500 ms, respectively. In each panel, S-shaped psychometric functions were observed for all subjects with either context $F0$. Also, the high-$F0$ context led to more Tone-2 responses for perceptually ambiguous target tones (i.e., those in the middle of the target tone series) than the low-$F0$ context. The rationalized arcsine transformed response percentages with contexts of a particular duration (i.e., the data in each panel) were analyzed by a separate two-way RM ANOVA with target onset $F0$ and context $F0$ as the two factors. When the contexts were 125 ms [Fig. 1(b)], tone recognition was significantly affected by both target onset $F0$ ($F_{8,80} = 93.37$, $p < 0.001$) and context $F0$ ($F_{1,10} = 22.60$, $p < 0.001$). However, the two factors did not significantly interact with each other ($F_{8,80} = 1.50$, $p = 0.17$). For the other context durations, the effects of target onset $F0$ and context $F0$, as well as their interactions were all significant. For the effect of target onset $F0$, $F_{8,80} = 183.12$, $p < 0.001$ in Fig. 1(c), $F_{8,80} = 115.76$, $p < 0.001$ in Fig. 1(d), $F_{8,80} = 149.74$, $p < 0.001$ in Fig. 1(e), and $F_{8,80} = 96.74$, $p < 0.001$ in Fig. 1(f). For the effect of context $F0$, $F_{1,10} = 65.34$, $p < 0.001$ in Fig. 1(c), $F_{1,10} = 36.53$, $p < 0.001$ in Fig. 1(d), $F_{1,10} = 35.67$, $p < 0.001$ in Fig. 1(e), and $F_{1,10} = 46.11$, $p < 0.001$ in Fig. 1(f). For the interactions between the two factors, $F_{8,80} = 9.65$, $p < 0.001$ in Fig. 1(c), $F_{8,80} = 5.19$, $p < 0.001$ in Fig. 1(d), $F_{8,80} = 11.20$, $p < 0.001$ in Fig. 1(e), and $F_{8,80} = 11.54$, $p < 0.001$ in Fig. 1(f). For each two-way RM ANOVA, *post hoc* Bonferroni t-tests were used to find the target tones that had significantly more Tone-2 responses with the high-$F0$ context than with the low-$F0$ context. Significantly different tone responses with the two contexts and the related p values are shown by asterisks in each panel of Fig. 1.

Each subject's tone recognition function with either the high- or low-$F0$ context of a particular duration was fit with the following two-parameter sigmoid function:

$$y = \frac{100}{1 + e^{-(x-x_0)/b}}, \tag{1}$$

where x_0 is the target onset $F0$ with 50% Tone-2 responses or the perceptual boundary between Tone 1 and Tone 2, and b is in inverse proportion to the slope of function and indicates a subject's sensitivity to the target pitch changes. Figure 2 shows the perceptual boundary with the high- or low-$F0$ context as a function of context duration. A two-way RM ANOVA found a significant effect of context $F0$ ($F_{1,10} = 41.02$, $p < 0.001$) but not of context duration ($F_{4,40} = 1.23$, $p = 0.31$) on the perceptual boundary. Also, the two factors significantly interacted with each other ($F_{4,40} = 9.71$, $p < 0.001$). *Post hoc* Bonferroni t-tests showed that the perceptual boundaries were significantly higher with the high-$F0$ context than with the low-$F0$ context for each of the context durations ($p = 0.04$ for 125 ms and $p < 0.001$ for 250 to 1500 ms). Perceptual boundaries with the high-$F0$ context did not significantly differ across the context durations ($p = 1.00$). In contrast, perceptual boundaries with the low-$F0$ context were significantly higher only for the 125-ms context duration than for any longer context duration ($p < 0.01$). According to another two-way RM ANOVA, the slope of function was also significantly affected by context $F0$ ($F_{1,10} = 8.74$, $p = 0.01$) but not by context duration ($F_{4,40} = 0.67$, $p = 0.62$). The two factors did not significantly interact with each other ($F_{4,40} = 1.25$, $p = 0.31$). Overall, the high-$F0$ context led to significantly steeper functions than the low-$F0$ context ($p = 0.01$).

4. Discussion

In this study, the target tones were more likely to be identified as Tone 2 by native Mandarin normal-hearing listeners with the high-$F0$ context than with the low-$F0$ context even when the contexts were as short as 125 ms. The contrastive effect of context $F0$ on tone recognition in terms of both response changes and boundary shifts saturated with the 250-ms contexts (longer contexts did not further enhance the context effect). The 250-ms contexts, which correspond to approximately a single syllable in a normal speaking rate, are thus long enough for listeners to extract the context pitch

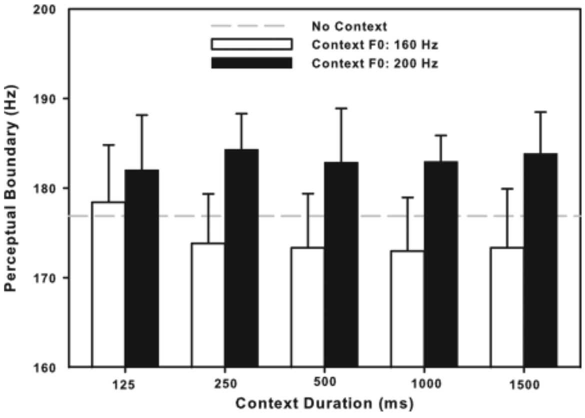

Fig. 2. Perceptual boundary between Tone 1 and Tone 2 with the high- (black bars) or low-$F0$ context (white bars) as a function of context duration. Symbols represent the mean, while error bars represent the standard deviation across subjects. The dashed gray line shows the mean perceptual boundary without context.

cues for Mandarin tone normalization. To more precisely locate the critical context duration for Mandarin tone normalization, more context durations around 250 ms should be tested, while the 1000- and 1500-ms context durations may not be necessary in future studies.

Consistent with previous studies of Mandarin tone normalization (e.g., Huang and Holt, 2009; Luo and Ashmore, 2014), this study again showed that Mandarin listeners made use of both intrinsic and extrinsic pitch cues to recognize Mandarin contour tones. The contrastive effect of context *F*0 on the recognition of target Tone 1–Tone 2 series was stronger in this study than in Huang and Holt (2009) and Luo and Ashmore (2014). For example, in Luo and Ashmore (2014), increasing the mean *F*0 of the preceding 1319-ms sentence from 160 to 200 Hz led to a 4-Hz increase in the mean Tone 1–Tone 2 boundary and 20% more Tone-2 responses for the target tone with the 175-Hz onset *F*0 (the largest response changes among the different target tones). In contrast, the mean Tone 1–Tone 2 boundary increased by 10 Hz and the target tone with the 175-Hz onset *F*0 had 40%–60% more Tone-2 responses when the context *F*0 was increased from 160 to 200 Hz for the preceding 500- or 1500-ms two-word phrase in this study (i.e., either the speaking rate or the duration of context was similar as in Luo and Ashmore, 2014). The Tone 1–Tone 2 boundaries for the same target tone series without context were lower in this study (∼177 Hz) than in Luo and Ashmore (2014; ∼181 Hz), suggesting that Mandarin listeners in this study were less sensitive to the intrinsic pitch changes within target tones. These listeners may thus rely more on extrinsic pitch cues to recognize target tones and exhibit stronger context effects than those in Luo and Ashmore (2014). Note that Mandarin listeners in this study had a similar linguistic experience as those in Luo and Ashmore (2014).

The response patterns across different context durations in this study suggest that the critical context duration for robust Mandarin tone normalization was 250 ms, which was shorter than the 450-ms critical context duration for Cantonese tone normalization in Eramela (2002). Note that both studies tested a similar set of context durations below 500 ms but found two different critical context durations for Mandarin and Cantonese tone normalization. The different results may be due to the different changes in context *F*0 but not the different languages tested in the two studies. Context effects are generally weaker for Mandarin contour tone recognition than for Cantonese level tone recognition, because there are more intrinsic pitch cues for Mandarin contour tone recognition (e.g., Wong and Diehl, 2003; Huang and Holt, 2009). This suggests that the contexts that are just long enough for Cantonese level tone normalization may not contain sufficient context pitch cues to affect Mandarin contour tone recognition. As such, Mandarin tone normalization should have a longer critical context duration than Cantonese tone normalization. Also, listeners may need a longer duration to extract the pitch information from the Mandarin contexts with different contour tones than from the Cantonese contexts with a single level tone. Instead, the critical context duration was shorter for tone normalization in Mandarin than in Cantonese, possibly because there was a larger *F*0 change between the Mandarin contexts in this study (3.9 semitones) than between the Cantonese contexts in Eramela (2002; 2 semitones). Future studies may use a within-subjects design to confirm if larger changes in context *F*0 would reduce the critical context duration for tone normalization.

The speaking rate was increased with shorter contexts but reduced with longer contexts. Regardless of the context *F*0, the Tone 1–Tone 2 boundaries did not change with context durations $\geq$ 250 ms. This suggests that subjects did not use the speaking rate cues in the context to adjust their recognition of Tone 1 and Tone 2. The results differed from those of Jongman and Moore (2000), which found a contrastive effect of speaking rates on the recognition of Tone 2 and Tone 3 (i.e., more Tone-2 responses with longer contexts or slower speaking rates). While Tone 2 and Tone 3 may be partially distinguished based on their durations (i.e., Tone 3 is typically longer than Tone 2), Tone 1 and Tone 2 have similar durations and their recognition does not rely on duration cues. That is why there was little speaking rate normalization for the recognition of Tone 1–Tone 2 series.

When the context duration was reduced to 125 ms in this study, the Tone 1–Tone 2 boundaries with the low-$F0$ context significantly increased, while those with the high-$F0$ context decreased slightly, but not significantly. Paired t-tests showed that subjects had similar Tone 1–Tone 2 boundaries with the 125-ms low-$F0$ context as without context ($t_{10} = 0.98$, $p = 0.35$), but had significantly higher Tone 1–Tone 2 boundaries with the 125-ms high-$F0$ context than without context ($t_{10} = 3.69$, $p = 0.004$; see Fig. 2). The 125-ms low-$F0$ context only had ~12 pitch periods, which may not be enough to robustly represent the lower context $F0$. Target tones with the 125-ms low-$F0$ context may thus have been perceived as if they were in isolation. In contrast, the 125-ms high-$F0$ context had 6 more pitch periods than the 125-ms low-$F0$ context. The additional pitch periods may have allowed subjects to at least partially perceive the higher context pitch, as shown by the similar Tone 1–Tone 2 boundaries with the 125-ms and longer high-$F0$ contexts.

The critical context duration for Mandarin tone normalization was obtained for normal-hearing listeners in this study. Recently, Luo *et al.* (2014) found that profoundly deaf people using a cochlear implant together with a hearing aid had similar context-dependent tone recognition as normal-hearing listeners. It would be of interest to test if cochlear implant users need longer critical context durations for tone normalization than normal-hearing listeners, due to their poorer access to pitch cues with the implant.

Acknowledgments

We thank all subjects for their participation in this study. Research was supported in part by NIH (Grant No. R21-DC011844). We thank Dr. Alex Francis for helpful discussions and for providing Eramela's thesis, which was co-supervised by him and Dr. Valter Ciocca.

References and links

Boersma, P. and Weenink, D. (**2012**). Praat: Doing phonetics by computer. Ver. 5.3.17. http://www.fon.hum.uva.nl/praat/ (Last viewed August 10, 2013).

Chao, Y. R. (**1948**). *Mandarin Primer* (Harvard University Press, Cambridge, MA).

Eramela, E. (**2002**). "The effect of precursor duration on tone normalization in Cantonese," Bachelor's dissertation, the University of Hong Kong, Hong Kong, China.

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

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

Jongman, A., and Moore, C. B. (**2000**). "The role of language experience in speaker and rate normalization processes," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Vol. 1, pp. 62–65.

Luo, X., and Ashmore, K. B. (**2014**). "The effect of language experience on perceptual normalization of Mandarin tones and non-speech pitch contours," J. Acoust. Soc. Am. **135**, 3585–3593.

Luo, X., Chang, Y.-P., Lin, C.-Y., and Chang, R. Y. (**2014**). "Contribution of bimodal hearing to lexical tone normalization in Mandarin-speaking cochlear implant users," Hear. Res. **312**, 1–8.

Luo, X., and Fu, Q.-J. (**2004**). "Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants," J. Acoust. Soc. Am. **116**, 3659–3667.

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

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res. **28**, 455–462.

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.

Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83.