MAY 14 2018

Effects of fundamental frequency contour on understanding Mandarin sentences in bimodal hearing simulations

Fei Chen; Jing Chen

J. Acoust. Soc. Am. 143, EL354–EL360 (2018)
https://doi.org/10.1121/1.5037720

Effects of fundamental frequency contour on understanding Mandarin sentences in bimodal hearing simulations

Fei Chen[a)]
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen, China
fchen@sustc.edu.cn

Jing Chen
Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
janechenjing@pku.edu.cn

Abstract: Fundamental frequency (F0) contour carries important information for understanding a tonal language. The present work assessed the effects of F0 contour on understanding Mandarin sentences in bimodal hearing simulations, including three conditions of acoustic-only, electric-only, and combined stimulations. Test stimuli were synthesized Mandarin sentences, each word with a normal, flat, or randomly assigned lexical tone, and presented to normal-hearing Mandarin-speaking listeners to recognize. Experimental results showed that changing F0 contour significantly affected the perception of Mandarin sentences under all conditions of acoustic-only, electric-only, and combined stimulations. The combined-stimulation advantage was only observed for test stimuli with the normal F0 contour.

[QJF]
Date Received: February 11, 2018 **Date Accepted:** April 25, 2018

1. Introduction

A cochlear implant (CI) is presently the only medical treatment for partially restoring hearing to patients with profound-to-severe hearing loss. While most CI users are able to perceive sound and understand speech, there are still many challenges for the present CI technology, e.g., understanding a tonal language (e.g., Luo and Fu, 2006; Li *et al.*, 2014; Chang *et al.*, 2016; Yang and Zeng, 2017), speech perception in noise (e.g., Chang *et al.*, 2006), etc. Many hearing-impaired patients preserve some degree of low-frequency (LF) residual hearing and largely suffer from hearing loss at high frequency. The combined-stimulation advantage of bimodal hearing refers to an improvement in speech recognition when electric stimulation in CI is supplemented by LF acoustic information (e.g., Micheyl and Oxenham, 2012). Recent studies have consistently shown that adding LF acoustic information may significantly improve CI speech perception, such as in noise (e.g., Chang *et al.*, 2006) and for a tonal language (e.g., Luo and Fu, 2006; Li *et al.*, 2014; Yang and Zeng, 2017). Carroll *et al.* (2011) showed that the fundamental frequency (F0) cue significantly contributed to the benefit received with combined electric-and-acoustic (E + A) hearing. They attributed this benefit to the modulation of the frequency rather than the amplitude component. Micheyl and Oxenham (2012) compared models of the combined-stimulation advantage for speech intelligibility, and suggested that this advantage was sometimes due to the noninteractive combination of two (i.e., electric and acoustic) information sources. Brown and Bacon (2009) studied the contributions of F0, the amplitude envelope, and the voicing of speech to the intelligibility of simulated E + A stimuli. They replaced LF speech with a tone that was modulated in frequency to track F0 of the speech and/or was modulated in amplitude to track the envelope of LF speech. Their results showed that significant perceptual benefit could be provided by a tone carrying F0 and amplitude envelope cues.

While a number of works examined the factors affecting English speech understanding in combined electric-and-acoustic stimulation, only a few studies have been

[a)]Author to whom correspondence should be addressed.

done for Chinese speech (e.g., Luo and Fu, 2006; Li *et al.*, 2014; Chang *et al.*, 2016; Yang and Zeng, 2017). With bimodal hearing simulations, Luo and Fu (2006) reported that for CI patients with residual acoustic hearing, preserving low-frequency acoustic information could improve Chinese speech recognition in noise. Li *et al.* (2014) measured Mandarin tone, vowel, and consonant recognition in quiet and noise with Mandarin-speaking bimodal (i.e., with a CI and a hearing aid, HA) listeners. They tested with the CI-only and CI + HA conditions, and found combined-stimulation advantage for tone recognition in noise and vowel recognition in quiet, but not for consonant recognition. Chang *et al.* (2016) found that in quiet, Chinese-speaking CI users' tone recognition significantly improved, but their vowel recognition did not significantly change with the use of a HA. The variable benefits of bimodal hearing to Mandarin speech recognition across studies might be due to the different degrees of residual acoustic hearing in individual subjects. Yang and Zeng (2017) recently measured bimodal benefits in Mandarin-speaking CI subjects who had contralateral residual acoustic hearing. They found that CI subjects produced a significant bimodal benefit in word recognition in noise, and suggested that consonants and tones accounted for the bimodal benefit.

In a tonal language such as Mandarin Chinese, words of the same syllable are distinguished by their lexical tones. There are four Mandarin lexical tones, i.e., Tone 1 (high level), Tone 2 (mid-rising), Tone 3 (mid-falling-rising), and Tone 4 (high-falling), based on the various patterns of the F0 contour (Howie, 1976; Chen and Loizou, 2011). F0 is defined as the lowest frequency of a periodic waveform, generated by the opening and closing of the vocal cords. F0 contour carries the most distinctive information for lexical tone identification, and temporal envelope and duration are secondary cues contributing to lexical tone recognition (e.g., Fu *et al.*, 1998). The importance of lexical tone contour for Mandarin sentence perception has been noted in various studies (e.g., Chen *et al.*, 2014; Wang *et al.*, 2013). For normal-hearing (NH) listeners, while lexical tone contours do not influence sentence understanding in quiet, they significantly impact Mandarin sentence intelligibility in noise (e.g., Chen *et al.*, 2014; Wang *et al.*, 2013).

Given the importance of lexical tone contour for Mandarin speech recognition, the present work aimed to assess the effects of F0 contour on understanding Mandarin sentences in bimodal hearing simulations. Specially, this work modified the lexical tone contour of each word in Mandarin sentences and examined how this change of lexical tone contour affected Mandarin sentence recognition in vocoder simulations, which were commonly used to mimic the speech coding processes employed in the electric-only hearing with a CI and the combined electric-and-acoustic hearing with a CI and a HA. This work hypothesized that F0 contour played an important role for understanding Mandarin sentences processed by the electric-only and combined stimulations, and manipulating F0 contour might affect the combined-stimulation advantage for Mandarin sentence recognition.

2. Methodology

2.1 Subjects and materials

This experiment involved 8 listeners (6 males and 2 females) with NH (pure-tone thresholds better than 20 dB hearing level at octave frequencies from 125 to 8000 Hz in both ears). All subjects were native speakers of Mandarin Chinese and were paid for their participations.

The speech materials consisted of sentences from the Mandarin Hearing in Noise Test (MHINT) corpus (Wong *et al.*, 2007). Each MHINT list consists of 10 sentences, and each sentence contains 10 words. The lexical tone contour of each sentence was manipulated to yield a normal tone (NT), a flat tone (FT), or a randomly assigned tone (RT) using the methods described in Chen *et al.* (2014). Briefly, the sentences in the three tone conditions were generated by a TTS synthesis engine (NeoSpeech, 2010) and normalized to the same root-mean-square (RMS) power across all tone conditions. The characters for each MHINT sentence were first presented as a string of Mandarin Pinyin (i.e., Mandarin phonetic symbol), e.g., "ta1 chuan1 le1 yi2 jian4 hui1 ge2 zi4 shang4 yi1" for the MHINT sentence "She wears a gray plaid jacket" in English, and digits 1, 2, 3, and 4 in Pinyin indicated the four lexical tones for each Mandarin word. The Pinyin strings were used as the input for the TTS engine to produce the synthesized MHINT sentences, and the lexical tone for each word could be flexibly altered by changing the tone value in each Pinyin. In the NT condition, the sentences were generated by using the original lexical tone for each Mandarin word of the MHINT sentences, and the sentence quality was similar to that produced by Mandarin speakers in their normal conversational styles. The F0 ranged from 180 to 420 Hz in the NT condition. In the FT condition, the TTS synthesis engine generated each word of the MHINT sentences in flat tone by setting each Pinyin in Tone 1. In the RT condition, each word of the MHINT sentences was synthesized with a Pinyin in a randomly selected tone (from Tones 1 to 4). More details on synthesizing the MHINT sentences in the three tone conditions can be found in Chen *et al.* (2014). Figure 1 exemplifies the F0 contours of a Mandarin sentence under the NT, FT and RT conditions synthesized by the TTS software. A steady-state speech-spectrum shaped noise was used to mask the MHINT sentences at 5 dB signal-to-noise-ratio (SNR), and the SNR level was chosen to avoid the ceiling/floor effects.

2.2 Signal processing

This experiment included three signal processing conditions. The first condition simulated the acoustic-only (A-only) stimulation. Speech signal was processed by low-pass (LP) filtering to generate the LP-processed stimuli. LP filtering was implemented by using a linear-phase finite impulse response filter with filter order of $10 \times$ fs/fcut, where fs is the sampling rate (16 kHz) and fcut is the LP cutoff frequency (fcut = 600 Hz in this study).

The second signal processing condition simulated an eight-band electric-only (E-only) stimulation by using a noise vocoder. To implement the noise vocoder, speech signals were first processed through a pre-emphasis filter (first-order high-pass filter with 1200-Hz cutoff frequency). Then, signals were bandpass-filtered into eight frequency bands between 80 and 6000 Hz with sixth-order Butterworth filters. Cutoff frequencies for the channel allocation of bandpass filters were (80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000 Hz), computed according to the cochlear frequency-position mapping function (Greenwood, 1990). From each band, the envelope was extracted by half-wave rectification and LP filtering with a 200-Hz cutoff frequency by way of a fourth-order Butterworth filter. White noise was used as the carrier signal, and amplitude-modulated by the extracted envelope. Output from each band was further band-limited with the same bandpass filter at that band. All amplitude-modulated noises (with band-limiting processing) were summed to generate the noise-vocoded stimulus, with its amplitude adjusted to have the same RMS power as the input speech signal.

The third signal processing condition simulated the combined electric-and-acoustic stimulation. To simulate the effect of combined electric-and-acoustic with residual hearing <600 Hz, we combined the LP stimulus described earlier with the upper five bands (ranging from 724 to 6000 Hz) of the eight-band noise vocoder. Note that the level of the LP stimulus was equalized to that of the E-only simulation. Again, RMS power scaling was performed with respect to the input speech signal.

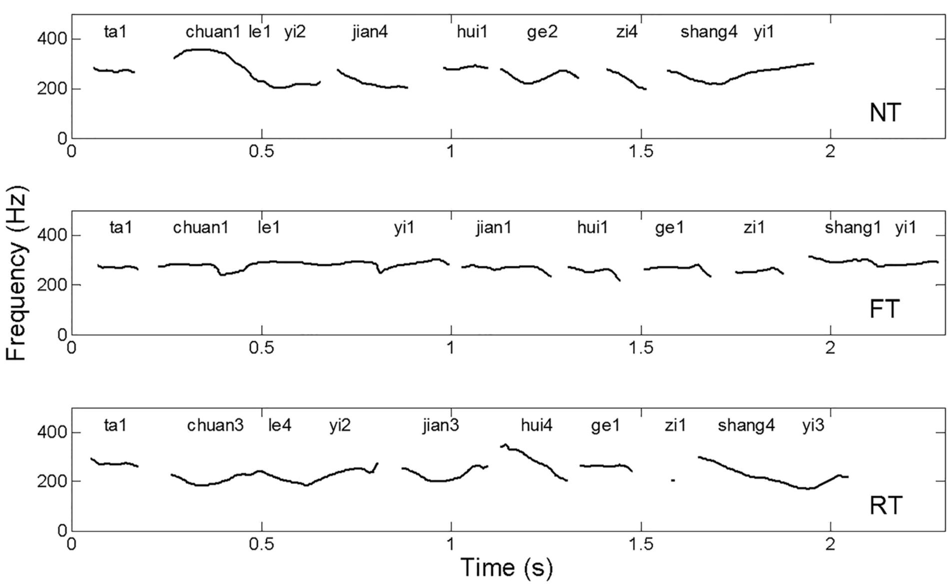

Fig. 1. The F0 contours of a Mandarin sentence (i.e., /da1 tʂʰuan1 1ɤ1 ji2 tɕiæn4 xuei1 kɤ2 tsi1 ʂɑŋ4 ji1/ or "She wears a gray plaid jacket" in English) under the NT, FT, and RT conditions synthesized by the TTS software. 1 = Tone 1, 2 = Tone 2, 3 = Tone 3, 4 = Tone 4 (Chen *et al.*, 2014).

2.3 Procedure

The experiment was performed in a sound booth, and stimuli were played to listeners monaurally through an HD 650 circumaural headphone (Sennheiser, Germany) set at a comfortable listening level. Before the actual testing session, each subject participated in a 10-min training session and was given four lists of 10 MHINT sentences (different with those used in testing session). The training session familiarized the subjects with the testing procedure and all different tested conditions. During the training session, the subjects were allowed to read transcriptions of the training sentences while they were listening to the sentences. In the testing session, the order of the conditions was randomized across subjects, and the subjects were asked to repeat orally all of the words they heard. In addition, the lists were randomized across listeners. The sentences used during testing were not the same as any of the training sentences. Each subject participated in a total of 18 conditions [= 2 SNR levels (i.e., quiet and 5 dB) × 3 tone conditions (i.e., FT, RT and NT) × 3 signal processing conditions (i.e., A-only, E-only and combined E + A)]. One list of 10 Mandarin sentences was used per tested condition, and none of the sentences was repeated across conditions. Subjects were allowed to listen to each stimulus a maximum of three times, and were asked to repeat as many words as they could recognize. A simple custom software interface was designed for the listening experiment, which each participant used to control the auditory delivery of the processed stimuli. During the testing session, a tester accompanied the participant and scored his/her response in the computer. A 5-min break was given every 30 min to avoid listening fatigue. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each MHINT list. The total testing time was one hour.

3. Results

Figure 2 shows the mean sentence recognition results for all conditions. Statistical significance was determined by using the percent recognition score as the dependent variable, and SNR level (i.e., quiet and 5 dB), tone condition (i.e., FT, RT and NT) and signal processing condition (i.e., A-only, E-only and combined E + A) as the three within-subject factors. Recognition scores were converted to rational arcsine units by using the rationalized arcsine transform (Studebaker, 1985). Three-way analysis of variance with repeated measures indicated significant effects of SNR level ($F_{1,7} = 160.01$, $p < 0.001$), tone condition ($F_{2,14} = 121.24$, $p < 0.001$), signal processing condition ($F_{2,14} = 407.59$, $p < 0.001$), a non-significant interaction between SNR level and tone condition ($F_{2,14} = 2.81$, $p = 0.09$), a significant interaction between SNR level and signal processing condition ($F_{2,14} = 24.41$, $p < 0.005$), a significant interaction between tone condition and signal processing condition ($F_{4,28} = 6.54$, $p < 0.05$), and a non-significant interaction among SNR level, tone condition and signal processing condition ($F_{4,28} = 1.27$, $p = 0.31$).

Post hoc pairwise comparisons were carried out at each group of conditions with the same SNR level and signal processing condition to analyze the effect of tone condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a p-value lower than 0.017 (= 0.05/3) were considered as significant. Results showed that at any group of conditions with the same SNR level and signal processing condition, the mean recognition scores under the NT condition were significantly ($p < 0.017$) larger than those under the FT or RT condition, and the mean

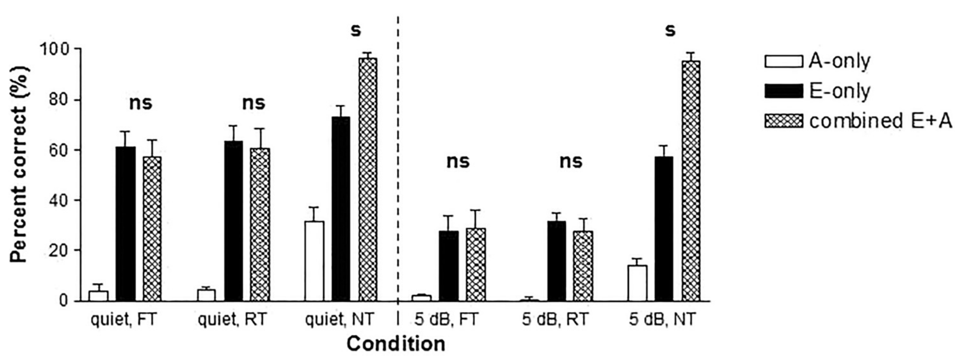

Fig. 2. Mean sentence recognition scores for all conditions. The error bars denote ±1 standard error of the mean. "s" and "ns" denote that the score difference between the E-only and combined E + A conditions is significant and non-significant, respectively.

recognition scores under the RT condition were non-significantly ($p > 0.017$) different with those under the FT condition.

Post hoc pairwise comparisons were carried out at each group of conditions with the same SNR level and tone condition to analyze the effect of signal processing condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a *p*-value lower than 0.017 (= 0.05/3) were considered as significant. Results showed that at any group of conditions with the same SNR level and tone condition, the mean recognition scores under the A-only condition were significantly ($p < 0.017$) smaller than those under the E-only or combined E + A condition. At any group of conditions with the same SNR level, the score differences between the E-only and E + A conditions were significant ($p < 0.017$) under the NT condition, but non-significant ($p > 0.017$) under the FT and RT conditions.

4. Discussion and conclusions

The present work assessed the effects of F0 contour on the understanding Mandarin sentences in bimodal hearing simulations. First, under acoustic-only simulation, this work showed that F0 contour carried a critical importance for understanding the low-pass filtered Mandarin sentences both in quiet and in noise. As seen in Fig. 2, with a normal F0 contour, the low-pass (600 Hz cutoff frequency) filtered Mandarin sentences are still intelligible, i.e., 31.5% in quiet and 14.0% in noise. However, when the normal F0 contour is replaced with a flat or randomly assigned lexical F0 contour, the intelligibility score of the low-pass filtered Mandarin sentences significantly declined (i.e., towards 0.0%), which demonstrates the impact of F0 contour on understanding low-pass filtered Mandarin sentences and is consistent with findings in prior work (e.g., Wang *et al.*, 2018).

Second, results in this work showed that F0 contour also carried critical importance for understanding vocoded (i.e., with the eight-band noise vocoder) Mandarin speech both in quiet and in noise. The vocoding process preserves multi-band temporal envelope waveforms, and discards the underlying fine-structure or phase information. The F0 variation or F0 contour information is largely removed by the vocoding process (e.g., Shannon *et al.*, 1995). However, temporal envelope co-varies with the lexical tone contour in Mandarin speech (e.g., Fu *et al.*, 1998), and temporal periodicity cue in amplitude modulation may also convey tonal information with CI simulations to some degree (Luo *et al.*, 2008). Hence, the preserved F0-contour information in temporal envelope may aid the perception of vocoded Mandarin sentences. In contrast, the FT and RT conditions do not preserve any useful F0-contour information in envelope, making their recognition scores lower than that under the NT condition, as shown in Fig. 2. The findings in this study are consistent with results from previous work (e.g., Chen *et al.*, 2014; Wang *et al.*, 2013). Early work showed that while the F0-contour information was important for Mandarin sentence perception, its perceptual contribution could be compensated by other cues (e.g., context information) in quiet. Hence, the loss of F0 contour information did not significantly influence Mandarin sentence understanding in quiet. However, under adverse conditions (e.g., in noise), F0 contour played an important role for Mandarin sentence recognition. In the present work, the degraded phonemic cues with vocoder processing may have limited the use of context information, making tonal cues more important for Chinese speech recognition. In this case, F0 contour is an important cue to aid Mandarin sentence understanding, which largely accounts for the improved sentence recognition performance under the NT condition. Similarly, Luo and Fu (2004) studied the importance of tonal cues to Chinese speech recognition with cochlear implant simulation. They used noise bands, fixed-rate pulse trains, and F0-controlled pulse trains as the carriers in 2- and 4-channel CI simulations, and found that flat pitch contours of fixed-rate pulse trains significantly degraded Chinese tone and sentence recognition but not vowel and consonant recognition with CI simulations.

Third, the present work found that F0 contour was an important factor for understanding Mandarin sentences processed by combined electric-and-acoustic stimulation. Figure 2 shows that both in quiet and in noise, the recognition scores under the NT condition are significantly larger than those under the FT or RT condition. This indicates the importance of preserving the normal F0 contour for understanding Mandarin sentences processed by combined stimulation. In addition, under both FT and RT conditions, which contain abnormal F0 contour information, combining acoustic and electric stimulations does not yield better speech recognition relative to electric-only stimulation; however, the FT condition provides combined-stimulation advantage in sentence recognition, i.e., 96.0% vs 73.0% in quiet and 95.0% vs 57.0% at

5 dB SNR level. This suggests the importance of preserving the normal F0 contour on receiving the advantage of combined stimulation.

Many early studies assessed the perceptual effect of adding low-frequency acoustic information to electric hearing, and revealed the bimodal benefits in both non-tonal (e.g., Brown and Bacon, 2009; Carroll *et al.*, 2011; Micheyl and Oxenham, 2012) and tonal (e.g., Luo and Fu, 2006; Li *et al.*, 2014; Chang *et al.*, 2016; Yang and Zeng, 2017) languages. As Mandarin Chinese differs with English in terms of the perceptual contribution of F0-contour or tonal information which is largely contained in the low-frequency region, the present work specifically examined the impact of preserving correct tonal information on bimodal benefits in Mandarin sentence perception. Results clearly indicated that preserving correct tonal information (i.e., in the NT condition) is important for achieving the observed bimodal benefits, which is the main novelty of the present work. In other words, presenting low-frequency acoustic information is not sufficient for receiving the bimodal benefits in Mandarin sentence perception, as bimodal benefits also require the correct tonal information presented. Nowadays, many speech processing approaches (e.g., noise suppression) are being employed in the E-only or bimodal hearing (e.g., Lai *et al.*, 2017). The implication of the present work is that the potential distortion to F0-contour or tonal information caused by speech processing might affect the bimodal benefits in understanding Mandarin sentences; and hence special caution needs be focused on controlling the level of distortion applied to tonal information.

In conclusion, the present work studied the effects of F0 contour on understanding Mandarin sentences in bimodal hearing simulations. Three tone conditions were tested in this work, including normal, flat and randomly assigned F0 contours. Results consistently showed that changing F0 contour significantly affected Mandarin sentence recognition under all tested conditions of acoustic-only, electric-only and combined stimulations. In addition, only the tested condition containing the normal F0 contour yielded the advantage of combined stimulation in understanding Mandarin sentences.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61571213, 61771023), and the Basic Research Foundation of Shenzhen (Grant No. JCYJ20170817110841907).

References and links

Brown, C. A., and Bacon, S. P. (**2009**). "Low-frequency speech cues and simulated electric-acoustic hearing," J. Acoust. Soc. Am. **125**, 1658–1665.

Carroll, J., Tiaden, S., and Zeng, F. G. (**2011**). "Fundamental frequency is critical to speech perception in noise in combined acoustic and electric hearing," J. Acoust. Soc. Am. **130**, 2054–2062.

Chang, J. E., Bai, J. Y., and Zeng, F. G. (**2006**). "Unintelligible low-frequency sound enhances simulated cochlear-implant speech recognition in noise," IEEE Trans. Biomed. Eng. **53**, 2598–2601.

Chang, Y., Chang, R. Y., Lin, C. Y., and Luo, X. (**2016**). "Mandarin tone and vowel recognition in cochlear implant users: Effects of talker variability and bimodal hearing," Ear Hear. **37**, 271–281.

Chen, F., and Loizou, P. C. (**2011**). "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," J. Acoust. Soc. Am. **129**, 3281–3290.

Chen, F., Wong, L. L. N., and Hu, Y. (**2014**). "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech Lang. Hear. Res. **57**, 338–345.

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

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, England).

Lai, Y. H., Chen, F., Wang, S. S., Lu, X. G., Tsao, Y., and Lee, C. H. (**2017**). "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implantation," IEEE Trans. Biomed. Eng. **64**, 1568–1578.

Li, Y., Zhang, G., Galvin, J. J., III, and Fu, Q. J. (**2014**). "Mandarin speech perception in combined electric and acoustic stimulation," Plos One **9**, e112471.

Luo, X., and Fu, Q. J. (**2004**). "Importance of pitch and periodicity to Chinese-speaking cochlear implant patients," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, Vol. 4, pp. 1–4.

Luo, X., and Fu, Q. J. (**2006**). "Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations," J. Acoust. Soc. Am. **120**, 2260–2266.

Luo, X., Fu, Q. J., Wei, C. G., and Cao, K. L. (**2008**). "Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users," Ear Hear. **29**, 957–970.

Micheyl, C., and Oxenham, A. J. (**2012**). "Comparing models of the combined-stimulation advantage for speech recognition," J. Acoust. Soc. Am. **131**, 3970–3980.

NeoSpeech (**2010**). version 3.11.1.1, http://www.neospeech.com (Last viewed April 19, 2018).

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

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

Wang, J. J., Shu, H., Zhang, L. J., Liu, Z. X., and Zhang, Y. (**2013**). "The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility," J. Acoust. Soc. Am. **134**, EL91–EL97.

Wang, L., Zheng, D. C., and Chen, F. (**2018**). "Understanding low-pass-filtered Mandarin sentences: Effects of fundamental frequency contour and single-channel noise suppression," J. Acoust. Soc. Am. **143**, EL141–EL146.

Wong, L. L. N., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear. **28**, 70S–74S.

Yang, H. I., and Zeng, F. G. (**2017**). "Bimodal benefits in Mandarin-speaking cochlear implant users with contralateral residual acoustic hearing," Int. J. Audiol. **56**, S17–S22.