Intelligibility of naturally produced and synthesized Mandarin speech by cochlear implant listeners

Ying Shi,¹ Jingyuan Chen,¹ Yue Gong,¹ Biao Chen,¹ Yongxin Li,^{1,a)} John J. Galvin III,² and Qian-Jie Fu³

¹Department of Otolaryngology, Head and Neck Surgery, Beijing TongRen Hospital, Capital Medical University, Beijing 100730, People's Republic of China

²House Ear Institute, Los Angeles, California 90057, USA

³Department of Head and Neck Surgery, David Geffen School of Medicine, UCLA, Los Angeles, California 90095, USA

(Received 11 December 2017; revised 13 April 2018; accepted 25 April 2018; published online 15 May 2018)

Mandarin is a tonal language, and it is important to preserve lexical tone information in synthesized speech. With natural speech, Chinese cochlear implant (CI) users have difficulty perceiving voice pitch cues important for lexical tone perception; it is unclear whether this difficulty persists in Mandarin synthesized speech. In this study, intelligibility of naturally produced and synthesized Mandarin speech was measured in Chinese CI listeners; intelligibility was also measured in a control group of normal-hearing (NH) listeners. Five synthesized voices were selected to represent different talker genders (male, female, child), speaking rates (normal, slow), and speaking styles (emotional, accent). The data showed that while modern Mandarin text-to-speech (TTS) systems can provide perfect speech intelligibility for NH listeners, overall intelligibility was much poorer for CI than for NH listeners. CI performance was significantly poorer with synthesized speech than with natural speech ($p < 0.001$). CI listeners were highly sensitive to the “extra-atypical” synthesized emotional and accented speech. Performance with each of the synthesized speech types was significantly correlated with performance with natural speech in CI users ($p < 0.01$ in all cases). While modern TTS systems offer educational and communication benefits to CI users and hearing-impaired individuals, the selection of synthesized voices should be carefully considered in education applications of TTS for hearing-impaired individuals, especially CI children, since poor intelligibility performance may affect language learning. © 2018 Acoustical Society of America.

<https://doi.org/10.1121/1.5037590>

[JFL]

Pages: 2886–2891

I. INTRODUCTION

The development of text-to-speech (TTS) systems has profoundly affected the field of augmentative and alternative communication (AAC) as the intelligibility and naturalness of TTS systems has been steadily improved over time. The availability of high-quality TTS systems can provide accessibility to people with communication difficulties. Some of the most important and useful TTS applications are reading and communication aids for the blind, deaf, and/or vocally handicapped individuals. TTS systems can also be used in many educational and interactive dialogue situations. Increasingly, synthesized speech is used for commercial communications. Indeed, synthesized speech is almost unavoidable for voice menu systems in telecommunications.

In a typical TTS system, the targeted speech segments are either retrieved from data inventory (concatenation approach)^{1,2} or synthesized using a source-filter model approach.^{3,4} The concatenation approach can provide more natural sounding speech but requires a large speech inventory. However, it is possible to achieve a high quality TTS system with a smaller speech inventory by using sub-word or variable-length unit concatenation.⁵

It is important that TTS systems have high intelligibility to be effective in AAC applications such as reading and communication aids. For normal-hearing (NH) listeners, many TTS systems offer high intelligibility and naturalness for a variety of speaking styles in native languages. Venkatagiri⁶ found that intelligibility of TTS systems was 22% less intelligible and nearly three times as many phoneme errors than for natural speech. Venkatagiri⁷ also found that intelligibility for non-native listeners was greatly reduced, especially with the concatenation and formant synthesis approach. Little is known about the intelligibility of TTS systems for hearing-impaired listeners such as cochlear implant (CI) users. Ji *et al.*⁸ evaluated the intelligibility of natural and synthesized English speech in native NH and CI listeners. For both subject groups, performance was significantly better with natural than with synthesized speech. However, the performance deficit with synthesized speech was much greater for CI listeners, especially with faster speaking rates. Due to the limited spectral resolution, CI users have greater difficulty perceiving “atypical” speech (i.e., speech that is different from the clear speech used in clinical evaluations) than do NH listeners, including conversational speech,⁹ speech with vocal emotion,^{10–12} accented speech from non-native talkers,¹³ and whispered speech.^{12,14} Depending on the accuracy or voices used in TTS systems,

^{a)}Electronic mail: entlyx@sina.com

CI users may also have difficulty with synthesized, atypical speech.

Mandarin Chinese is a tonal language in which lexical tones convey linguistic meaning. Fundamental frequency (F0) is the primary cue for lexical tone perception, along with duration and amplitude cues that co-vary with F0.^{15,16} The four major tonal patterns are Tone 1 (high flat), Tone 2 (rising), Tone 3 (falling-rising), and Tone 4 (falling). Lexical tone perception contributes strongly to Mandarin sentence recognition.¹⁷ It is a challenge for TTS systems to encode lexical tone information alongside other prosodic cues in speech.

As noted above, CI users have greater difficulty with atypical speech, especially where voice pitch cues are important. Depending on the accuracy of the TTS system, CI users also have greater difficulty with synthesized Mandarin, as distortions to lexical tone cues might further degrade intelligibility, relative to natural speech. Given the increased use of TTS in educational and commercial settings, it is important to understand CI users' perception of synthesized speech. In this study, the intelligibility of natural and synthesized Mandarin speech was evaluated in native Chinese NH and CI listeners; TTS speech was produced for a variety of voice genders, speaking rates, and speaking styles.

II. METHODS

A. Subjects

Seventeen Mandarin-speaking Chinese CI users also participated in the study (10 males and 7 females). The mean age at testing was 23.7 yr (range = 15–64 yr), and the mean CI experience was 8.3 yr (range = 1.4–16.6 yr). Note that for some subjects, age at hearing level (HL) diagnosis is <1 yr; this rough approximation is because consistent screening was not available to these newborns, but rather indicated according to parent reports. Seventeen NH subjects (8 males

and 9 females; mean age = 26.6 yr, range = 15–63 yr) also participated in the study as a control group. All NH subjects had pure tone thresholds <20 dB HL at all audiometric frequencies between 125 and 6000 Hz. There was no significant difference in age at testing between these the CI and NH groups ($p=0.515$). CI subject demographic information is shown in Table I. In compliance with ethical standards for human subjects, written informed consent was obtained from all participants before proceeding with any of the study procedures. This study was approved by Institutional Review Board in Beijing Tongren Hospital, Capital Medical University.

B. Test materials and procedures

Sentence recognition was measured in quiet using Mandarin speech perception (MSP) sentences.^{18,19} The validated MSP stimulus set consists of eight lists (10 sentences in each list) of easy difficulty. Each sentence includes seven monosyllabic words. Table II shows an example MSP test list. All sentences were naturally produced by a single female talker in slow clear speaking style (*F-Nat*) or were synthesized using the iFlyTek TTS engine (www.iflytek.com). The iFlyTek TTS engine utilized a “corpus-based” approach that can synthesize very natural and high intelligible sounds with a large size of the corpus. Five synthesized voices were selected to represent different talker genders, speaking rates, and speaking styles, including:

- Voice #1 (*F-Slow*): The voice was synthesized with the corpus spoken by a female talker in broadcasting style. The voice was synthesized with slow speaking rate, see below for the detailed definition about speaking rate.
- Voice #2 (*F-Norm*): The voice was synthesized with the corpus spoken by the same female talker as Voice #1 except that the voice was synthesized with normal speaking rate.

TABLE I. CI subject demographic information. Coch = Cochlear Corp. device; AB = Advanced Bionics device. For years of HA use, Ipsilateral refers to HA in the CI ear, contralateral refers to HA in the contralateral ear to the CI.

Subject	Gender	Age at hearing loss (yr)	Age at test (yr)	CI exp (yr)	CI ear	Etiology	Device	Current HA use?	Years of HA use	MSP % correct
S1	M	<1	15	12.9	L	Unknown	Coch CI24C	No	0	100
S2	F	<1	15	11.4	L	LVAS	Coch CI24RCS	No	0	99
S3	M	<1	15	8.5	R	LVAS	AB HR90K	No	3 bilateral	76
S4	M	2	16	14.3	L	Unknown	Coch CI512	No	0	70
S5	F	12	17	4.0	R	Unknown	AB HR90K	No	0	89
S6	F	<1	17	5.1	L	LVAS	AB HR90K	No	3 bilateral	69
S7	M	<1	18	6.0	L	LVAS	Coch CI24R	HA	11 bilateral	61
S8	M	<1	18	7.1	L	LVAS	Med-EL, C40+	No	1.5 bilateral	9
S9	M	2	19	16.1	R	Unknown	Coch CI24C	No	0	54
S10	F	2	19	15.9	R	Unknown	Coch CI512	No	0	69
S11	F	<1	20	1.4	R	LVAS	Med-EL Sonata	No	15 Ipsilateral	97
S12	M	<1	21	4.0	L	LVAS	AB HR90K	No	17 bilateral	97
S13	F	4	22	16.6	R	Unknown	Med-EL Sonata	No	0	97
S14	M	<1	23	3.6	R	LVAS	AB HR90K	No	0	100
S15	M	31	38	5.7	L	Trauma	Med-EL, C40+	No	0	60
S16	M	43	48	3.0	L	Temporal bone fracture	AB HR90K	No	0	79
S17	F	57	64	5.3	R	Unknown	AB HR90K	No	1 Contralateral	57

TABLE II. Example of sentence test list.

#	Chinese Character	Chinese Pinyin	English Translation
1	今天的阳光真好	jīn tiān de yáng guāng zhēn hǎo	It's a nice sunny day.
2	节假日不用门票	jié jiǎ rì bù yòng mén piào	No ticket is needed during holiday.
3	晚上一块去跳舞	wǎn shàng yī kuài qù tiào wǔ	Let's go dancing tonight.
4	对面有两所高中	duì miàn yǒu liǎng suǒ gāo zhōng	There are two high schools across street.
5	这些衣服洗过吗	zhè xiē yī fú xǐ guò ma	Have these clothes been washed yet?
6	北京近来很寒冷	běi jīng jìn lái hěn hán lěng	It's very cold in Beijing recently.
7	他家每年放鞭炮	tā jiā měi nián fàng biān pào	He touches off the firecracker every year.
8	外孙出生在农村	wài sūn chū shēng zài nóng cūn	Grandson was born in rural areas.
9	星期二别打篮球	xīng qī èr bié dǎ lán qiú	Don't play basketball on Tuesday.
10	短裙长度正合适	duǎn qún cháng dù zhèng hé shì	The length of short skirt is appropriate.

- Voice #3 (**M-Norm**): The voice was synthesized with the corpus spoken by a male talker in broadcasting style. The voice was synthesized with normal speaking rate.
- Voice #4 (**F-Emot**): The voice was synthesized with the corpus spoken by a female talker in storytelling style with emotional quality. The voice was synthesized with normal speaking rate.
- Voice #5 (**B-Acc**): The voice was synthesized with the corpus spoken by a five-year-old boy in the accented style of Shin-chan, a main character in the popular Japanese manga and anime series “Crayon Shin-chan.” The selection of this voice was due to its uniqueness of its accents and the popularity in the young Chinese viewers. The voice was synthesized with normal speaking rate.

The TTS engine allows for independent adjustment of speaking rate and pitch. The range of parametric variations allows the adjustment of both speed and pitch. The range of manipulation is 0 to 100. For example, the fundamental frequency (F0) of speech synthesized with the 100 setting is twice as high as with the 50 setting while the speaking rate of speech synthesized with the 100 setting is twice as fast as with the 0 setting. For all synthesized speech, the pitch and speaking rate were set to 50, except for the F-Slow condition in which speaking rate was set to 0. Detailed information regarding F0 and speaking rate for the recorded and synthesized stimuli are shown in Table III. Note that the speaking rate was similar between F-Nat (natural speech from the MSP materials) and F-Slow; during recording of the MSP sentences, the talker was instructed to speak slowly and clearly. All stimuli were normalized to have the same long-term RMS amplitude (65 dBA). Please see Mm. 1, which shows the different voices of a sentence as an example.

Mm. 1. Different voices of a sentence. This is a file of type “wav” (522 KB).

Figure 1 shows spectrograms and electrograms (CI electrode stimulation patterns over time) for an example MSP sentence (Sentence #3 from the list shown in Table II)

produced for each stimulus type. In the spectrograms, slight distortions to F0 contours can be observed between the F-Nat and the F-Slow, F-Norm, and M-Norm stimuli, with even greater distortions for the F-Emot and B-Acc stimuli. The electrograms show the changes in the electrode stimulation patterns associated with the F0 contours are a potential cue for tone recognition for the F-Nat, F-Slow, F-Norm, and M-Norm, but change less for the F-Emot and B-Acc stimuli. The electrograms illustrate potential problems for some TTS voices, as spectral envelope cues may be greatly distorted.

All CI subjects were tested with their clinical processors and settings; these were not changed during the course of testing. For all subjects, stimuli were presented in the sound field at 65 dBA via a single loudspeaker; subjects were seated directly facing the loudspeaker at a 1-m distance. During testing, a sentence list was randomly selected, and sentences were randomly selected from within the list (without replacement) and presented to the subject, who repeated each sentence as accurately as possible. The experimenter calculated the percent of words correctly identified in sentences. All words in the sentences were scored, resulting in a total of 70 keywords for each list. Each keyword contains three components (initial consonant, final vowel, and tone). The response was scored as correct only if the subject correctly identified the keyword (and by association, the three components of initial consonant, final vowel, and tone). All test materials were presented once (without repeat), and no feedback was provided regarding the correctness of the response. The test order for the different speaking styles was randomized within and across subjects. Subjects were allowed to take breaks any time they felt fatigued.

III. RESULTS

Figure 2 shows boxplots of sentence recognition scores for CI subjects for the different speech types. Table IV shows CI subjects' individual recognition scores with the different speech types, and Table V shows the mean scores

TABLE III. Mean F0 (Hz) and speaking rate (words/second) across all sentences within each speech stimulus type.

	F-Nat	F-Slow	F-Norm	M-Norm	F-Emot	B-acc
F0 (std)	223 (14.8)	265 (17.0)	267 (16.4)	127 (9.8)	333 (13.8)	284 (5.9)
Rate (std)	3.56 (0.24)	3.41 (0.31)	4.51 (0.30)	4.79 (0.38)	4.07 (0.23)	3.04 (0.16)

FIG. 1. (Color online) Spectrograms and electroencephalograms for example speech stimuli. The Chinese, Pinyin, and English translations are shown at top. The numbers above each stimulus show the lexical tone for each syllable. The electroencephalograms were generated using similar parameters as for the default clinical parameters for the ACE signal processing in Cochlear Corp devices.

and standard deviation for CI subjects and the NH control group for the different speech types. CI subjects performed best with F-Nat, with poorer performance for the F-Slow, F-Norm, and M-Norm stimuli, and even poorer performance with the F-Emot and B-Acc stimuli. Note that there was also a wide range of CI performance; some subjects performed near NH levels with the F-Nat, F-Slow, F-Norm, and M-Norm stimuli, while others performing 50% correct or worse with the synthesized speech. As shown in Table V, most NH subjects performed near ceiling (100% correct) with both natural and synthesized speech.

Because of the large performance difference between CI and NH subjects, data were analyzed separately for the two subject groups using one-way repeated-measures analyses of variance (RM ANOVAs), with speech type (F-Nat, F-Slow, F-Norm, M-Norm, F-Emot, B-Acc). For CI subjects, there was a significant effect of speech type [$F(5,80)=53.8$, $p<0.001$; $\eta^2=0.771$]. *Post hoc* Bonferroni pair-wise comparisons showed that performance was significantly better with F-Nat than with any of the synthesized voices ($p<0.001$ in all cases), significantly better with F-Norm than with F-Emot, or B-Acc ($p<0.05$ in both cases), significantly better with F-Slow than with F-Emot or B-Acc

($p<0.05$ in both cases), significantly better with M-Norm than with B-Acc or F-Emot ($p<0.05$ in both cases), and significantly better with F-Emot than with B-Acc ($p<0.05$). For NH subjects, one-way RM analysis of variance (ANOVA) on ranks showed a slightly significant effect of speech type [$\chi^2(5)=13.18$, $p=0.022$]. However, *Post hoc* Tukey pair-wise comparisons showed no significant difference between any two speech types.

Each panel of Fig. 3 shows individual CI subjects' performance with the different synthesized speech types as a function of performance with the naturally uttered MSP sentences (F-Nat). Pearson correlation analyses were performed on the data shown in Fig. 3; r and p values are shown in each panel. After Bonferroni adjustment for multiple comparisons (adjusted $p=0.01$), significant correlations were observed between each of the synthesized speech types and the F-Nat. Demographic variables age at testing and CI experience were also compared to performance with the different speech types. There were no significant correlations among any of the demographic variables and speech types ($p>0.05$ in all cases).

FIG. 2. Boxplots of sentence recognition scores for the different speech stimuli for CI subjects. The boxes show the 25th and 75th percentiles, the error bars show the 5th and 95th percentiles, the symbols show outliers, the solid horizontal line shows the median and the dashed horizontal line shows the mean.

TABLE IV. Individual CI subject recognition scores (in percent correct) for different speech types.

Subject	F-Nat	F-Slow	F-Norm	M-Norm	F-Emot	B-Acc
S1	97	76	89	61	56	47
S2	100	100	93	66	91	30
S3	54	54	48	29	10	16
S4	69	53	64	60	44	19
S5	99	87	91	70	57	39
S6	76	54	30	26	24	1
S7	97	85	84	67	64	47
S8	97	62	65	73	64	24
S9	100	89	94	87	70	49
S10	70	51	58	69	17	7
S11	89	54	54	83	34	13
S12	60	29	40	16	23	7
S13	69	59	64	69	40	31
S14	79	66	34	46	34	9
S15	97	93	84	86	83	49
S16	57	50	64	60	40	0
S17	61	38	46	57	56	29

TABLE V. Mean performance and standard deviation for different speech types for CI and NH subject groups.

	F-Nat	F-Slow	F-Norm	M-Norm	F-Emot	B-acc
CI (std)	80.58 (19.85)	64.63 (19.85)	64.84 (21.40)	60.16 (21.21)	47.56 (23.60)	24.45 (17.70)
NH (std)	99.87 (0.42)	99.87 (0.42)	99.87 (0.42)	98.17 (2.40)	98.84 (2.19)	99.09 (1.17)

IV. DISCUSSION

The present data demonstrated that while modern Mandarin TTS systems such as iFlyTek can provide excellent intelligibility for NH listeners, intelligibility was much poorer for CI listeners. CI performance with synthesized speech was significantly poorer than with natural speech, consistent with previous studies.⁸ CI listeners were highly sensitive to the “extra-atypical” synthesized emotional and accented speech. CI performance was also highly variable, ranging from 54%–100% even with the recorded MSP sentences. Seven out of 17 CI subjects performed nearly perfectly (97%–100% correct for the recorded MSP sentences (F-Nat). However, mean performance for these subjects dropped to 40.6% for accented synthesized speech (B-Acc).

There was no significant difference between the F-Slow and F-Norm stimuli, similar to finding by Su *et al.*¹² for naturally produced speech at slow and normal speaking rates. This suggests that slowing down clear or synthetic speech might not necessarily improve intelligibility, relative to normal rates. As shown by Ji *et al.*,⁸ slowing down synthetic speech from fast to normal rates may improve intelligibility.

Different from Su *et al.*,¹² who found no significant difference between naturally produced clear speech and emotional speech, CI performance with F-Emot was significantly poorer than with F-Slow, F-Norm, or M-Norm. It is possible that the speech synthesis used to produce the F-Emot stimuli did not preserve key speech features such as lexical tone F0, amplitude, and/or duration. Mean F0 was highest for F-Emot (Table II), and F0 contours for Tones 3 and 4 appear to have

a narrower frequency range and a different harmonic distribution, compared to the F-Slow, F-Norm, or M-Norm stimuli (Fig. 1). For B-Acc, the F0 contours and harmonic distribution are similarly distorted, with additional distortions to the syllable amplitude (note that B-Acc had the slowest speaking rate of all stimuli). Note that CI performance with B-Acc was poorer than for whispered speech in Su *et al.*,¹² where voice pitch cues were unavailable. Despite the poor coding of F0 information with their device, CI listeners were especially susceptible to these distortions to voice pitch and/or spectral envelope cues in the synthesized speech. Data from Su *et al.*¹² suggest that CI listeners are less susceptible to changes in F0 patterns when speech is naturally produced with emotion.

While modern TTS systems can provide excellent intelligibility for NH listeners, they may be less than ideal for CI patients. Optimization of speech synthesis may be quite different for CI users than for NH listeners. Different synthesis approaches or even different voices may produce different intelligibility in CI users. It may also be true that no single approach or voice may be best for all CI users. In this study, nearly half the subjects performed better with the M-Norm than with F-Norm, and the performance difference between the two speech types was as large as 29.7 percentage points. As such, it may be that some synthetic voices may be better for some CI listeners than others, depending on the relationship between the acoustic input and patient-related factors such as frequency allocation, electrode–neural interface, etc. Optimization for these patients may require selecting different synthesized voices. One limitation in this study was that

FIG. 3. Speech performance with different synthesized speech types as a function of naturally uttered speech (F-Nat, from the MSP sentences). Linear regressions, correlation coefficients, and p-values are shown in each panel.

only one male and female synthetic voice was tested with the normal rate. Alternate voices produced by different speech synthesis engines might produce different results. Also, the naturalness of speech was not tested in this study, and performance was acutely measured. It is possible that after longer experience with the different voices, intelligibility and naturalness might improve.

TTS systems offer potential educational and communication benefits to CI users and hearing-impaired individuals. There were significant correlations between CI users' performance with naturally uttered speech (F-Nat) and each of the synthesized speech types (Fig. 3). While absolute performance levels may have been poorer, the strong correlations suggest that synthesized speech may be a useful proxy for natural speech, especially within an educational context. TTS systems have even been legitimized as a "real" human voice (e.g., Stephen Hawking). Additional research is needed to optimize speech synthesis parameters for conditions where the spectral resolution is reduced, as in the CI case. The selection of synthesized voices should be carefully considered in education applications of TTS for hearing-impaired individuals, especially CI children, since poor intelligibility performance may affect language learning. The wide range of voices readily accessible with TTS may also provide useful training materials for CI users. Training with varied and/or difficult stimuli has shown to generally benefit training outcomes.²⁰ Such training materials could be easily developed and the level of difficulty could be effectively controlled by modifying the speaking rate and/or speaking style of the synthesized voices. Further studies are needed to determine whether training with synthesized speech generalizes to better understanding of naturally produced speech.

ACKNOWLEDGMENTS

We thank the subjects for their participation in this study. We also thank two anonymous reviewers for helpful comments. This work was partly supported by the National Institutes of Health (Grant No. R01-004792).

¹E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.* **9**, 453–468 (1990).

- ²R. Sproat, J. Hirschberg, and D. Yarowsky, "A corpus-based synthesizer," in *ICSLP 1992*, Alberta, Canada (1992), pp. 563–566.
- ³J. Allen, M. S. Hunnicutt, and D. Klatt, *From Text to Speech—The MITalk System* (MIT Press, Cambridge, MA, 1987).
- ⁴D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995 (1980).
- ⁵Z. Yu, D. Yue, Y. Zu, and G. Chen, "Word intelligibility testing and TTS system improvement," in *ICSP Proceedings* (2010).
- ⁶H. S. Venkatagiri, "Segmental intelligibility of four currently used text-to-speech synthesis methods," *J. Acoust. Soc. Am.* **113**(4), 2095–2104 (2003).
- ⁷H. S. Venkatagiri, "Phoneme intelligibility of four text-to-speech products to nonnative speakers of English in noise," *Int. J. Speech Technol.* **8**(4), 313–321 (2005).
- ⁸C. Ji, J. J. Galvin III, A. Xu, and Q.-J. Fu, "Effect of speaking rate on recognition of synthetic and natural speech by normal-hearing and cochlear implant listeners," *Ear Hear.* **34**, 313–323 (2013).
- ⁹S. Liu, E. del Rio, A. R. Bradlow, and F. G. Zeng, "Clear speech perception in acoustic and electric hearing," *J. Acoust. Soc. Am.* **116**(4), 2374–2383 (2004).
- ¹⁰X. Luo, Q.-J. Fu, and J. J. Galvin III, "Vocal emotion recognition by normal-hearing listeners and cochlear implant users," *Trends Amplif.* **11**, 301–315 (2007).
- ¹¹M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, A. M. Kulkarni, and J. A. Christensen, "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers," *Hear. Res.* **322**, 151–162 (2015).
- ¹²Q. Su, J. J. Galvin, G. Zhang, Y. Li, and Q. J. Fu, "Effects of within-talker variability on speech intelligibility in Mandarin-speaking adult and pediatric cochlear implant patients," *Trends Hear.* **20**, 1–16 (2016).
- ¹³C. Ji, J. J. Galvin, Y. P. Chang, A. Xu, and Q. J. Fu, "Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants," *J. Speech Lang. Hear. Res.* **57**(2), 532–554 (2014).
- ¹⁴Y. Li, G. Zhang, H. Y. Kang, S. Liu, D. Han, and Q. J. Fu, "Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users," *J. Acoust. Soc. Am.* **129**(6), EL242–EL247 (2011).
- ¹⁵Z. A. Liang, "The auditory perception of Mandarin tones," *Acta Phys. Sin.* **26**, 85–91 (1963).
- ¹⁶M. C. Lin, "The acoustic characteristics and perceptual cues of tones in standard Chinese," *Chin. Yuwen.* **204**, 182–193 (1988).
- ¹⁷Q. J. Fu, F. G. Zeng, R. V. Shannon, and S. D. Soli, "Importance of tonal envelope cues in Chinese speech recognition," *J. Acoust. Soc. Am.* **104**(1), 505–510 (1998).
- ¹⁸Q. J. Fu, M. Zhu, and X. Wang, "Development and validation of the Mandarin speech perception test," *J. Acoust. Soc. Am.* **129**(6), EL267–EL273 (2011).
- ¹⁹Y. Li, S. Wang, Q. Su, J. J. Galvin, and Q. J. Fu, "Validation of list equivalency for Mandarin speech materials to use with cochlear implant listeners," *Int. J. Audiol.* **56**(2), S31–S40 (2017).
- ²⁰S. Amitay, D. J. Hawkey, and D. R. Moore, "Auditory frequency discrimination learning is affected by stimulus variability," *Percept. Psychophys.* **67**(4), 691–698 (2005).