Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users

Yongxin Li and Guoping Zhang

*Department of Otolaryngology, Head and Neck Surgery, Beijing TongRen Hospital,
Capital Medical University, Beijing, P. R. China 100730
lyxent@hotmail.com, cmu-zgp@163.com*

Hou-yong Kang

*Department of Otorhinolaryngology, Head and Neck Surgery, The First Affiliated Hospital,
Chongqing Medical University, 1 Youyi Road, Yuzhong District, Chongqing, China 400016
kanghouyong@sohu.com*

Sha Liu and Deming Han

*Department of Otolaryngology, Head and Neck Surgery, Beijing TongRen Hospital,
Capital Medical University, Beijing, P. R. China 100730
liusha_ls@sohu.com, handm@trhos.com*

Qian-Jie Fu^{a)}

*Communication and Neuroscience Division, House Ear Institute,
2100 W. 3rd St., Los Angeles, California 90057
qfu@hei.org*

Abstract: Cochlear implant (CI) users' speech understanding may be influenced by different speaking styles. In this study, speech recognition was measured in Mandarin-speaking CI and normal-hearing (NH) subjects for sentences produced according to four styles: slow, normal, fast, and whispered. CI subjects were tested using their clinical processors; NH subjects were tested while listening to a four-channel CI simulation. Performance gradually worsened with increasing speaking rate and was much poorer with whispered speech. CI performance was generally similar to NH performance with the four-channel simulation. Results suggest that some speaking styles, especially whispering, may negatively affect Mandarin-speaking CI users' speech understanding.

© 2011 Acoustical Society of America

PACS numbers: 43.71.Bp, 43.71.Es, 43.71.Ky, 43.64.Me [DO]

Date Received: March 09, 2011 **Date Accepted:** March 31, 2011

1. Introduction

In general, normal-hearing (NH) listeners are able to understand speech regardless of variations in talker, speaking rate, and language context (e.g., [Sommers *et al.*, 1992](#); [Eskenazi, 1993](#)). [Eskenazi \(1993\)](#) identified four types of speech variation: (1) Voice quality (breathy, creaky, lax, whispery, tense, etc.), (2) speaking rate (slow, fast, etc.), (3) speaking style (e.g., clear versus conversational), and (4) task- or context-specific (newscasts, sports, concerts, professional, interview, etc.). Automatic speech recognition (ASR) studies have evaluated the effects of speaking styles on speech intelligibility (e.g., [Woodland, 2001](#)). In general, natural and realistic speech production reduced ASR intelligibility compared to clearly produced speech, even after incorporating speaker adaptation techniques. Previous studies with hearing-impaired listeners (e.g., [Picheny *et al.*, 1985](#); [Uchanski *et al.*, 1996](#)) and cochlear implant (CI) users ([Liu *et al.*, 2004](#)) have shown that clear speech is more intelligible than conversational speech.

^{a)}Author to whom correspondence should be addressed.

CI users are susceptible to the acoustic variability in speech signals, due to the limited spectral resolution of the implant device (Shannon *et al.*, 2004). For example, CI users exhibit significant talker variability effects (Chang and Fu, 2006). CI users may also perform better with one talker than another (Liu *et al.*, 2008). Most CI speech studies have used clear speech samples in quiet or noise. Little is known about the effects of speaking rate or voice quality on CI users' speech understanding.

Mandarin Chinese is a tonal language, and perception of lexical tones depends strongly on fundamental frequency (F0) cues (e.g., Lin, 1988). When F0 cues are unavailable, CI listeners may use amplitude envelope (2–50 Hz) and/or periodicity cues (50–500 Hz) to identify lexical tones (Fu *et al.*, 1998). Speech rate may also affect CI listeners' access to envelope and periodicity cues. Liang (1963) found that Mandarin-speaking NH listeners correctly understood 64.0% of whispered lexical tones, i.e., after F0 and harmonic fine structure cues were removed. Besides the absence of F0 cues, whispering may otherwise reduce the intelligibility of speech. Due to the importance of lexical tones to Mandarin Chinese speech understanding, Mandarin-speaking CI users may have great difficulty with whispered speech. Currently, there are no published reports regarding CI users' perception of whispered speech or speech produced with different speaking rates. In this study, Mandarin-speaking CI users' sentence recognition was evaluated for four speaking styles: slow speaking rate, normal speaking rate, fast speaking rate, and whispered speech.

2. Methods

Thirteen Mandarin-speaking, postlingually deafened, adult CI patients (eight male and five female, aged 21–56 years old) and eight NH listeners (four male and four female, aged 22–42 years old) participated in this study. Among the CI subjects, three were implanted with an Advanced Bionics device, five were implanted with a Med-El device, and five were implanted with a Cochlear device. Twelve of the thirteen CI subjects had more than one year of experience with their device; the remaining subject had six months of experience at the time of testing. For nine of the CI subjects, the etiology of deafness was unknown. For the remaining four CI subjects, the etiology of deafness was ototoxicity, large vestibular aqueduct syndrome, otitis media, or noise-induced. All subjects were paid for their participation, and all provided informed consent in accordance with the local Institutional Review Board.

Sentence recognition was measured in quiet using Mandarin Easy Sentence Test (MEST; developed at the House Ear Institute). The MEST materials consist of ten lists of ten sentences each. Each sentence includes seven monosyllabic words and is of easy difficulty. Both phonetic balancing (across lists) and word familiarity were carefully considered in the development of the test materials. List equivalency (in quiet) was confirmed in NH subjects listening to four-channel acoustic simulation of CI processing (see the following), with mean scores ranging from 91% to 94% correct across the ten lists.

All sentences were produced by a single female talker in four different speaking styles (slow, normal, fast, and whispered). At the time of recording, the talker had more than 10 years of professional experience as a broadcaster in a radio station. During the recording of sentence materials, the talker was instructed to produce sentences in a slow, normal, or fast manner, or for whispered speech, to produce speech, while whispering in a normal speaking rate. Each sentence was recorded several times and the sentence that best represented the targeted speaking style was used in the test materials for the present study. Table I shows the detailed information for the MEST test materials.

NH subjects were tested while listening to unprocessed speech as well as to a four-channel, sine-wave vocoded acoustic simulation of CI speech processing. A sine-wave vocoder was used instead of noise-band vocoder because our recent studies suggest that sine-wave vocoders better represent CI-mediated listening for pitch related tasks such as voice gender recognition (Fu *et al.*, 2005). CI users' voice gender

Table 1. Mean duration per sentence, words per sentence, and words per second for the experimental speech materials.

Test materials	Mean duration per sentence (ms)	Mean F0 (Hz)	Mean words per second
MEST slow rate	2826	221.37	2.48
MEST normal rate	1870	229.55	3.74
MEST fast rate	1234	239.57	5.67
MEST whispered	1968	N/A	3.56

discrimination depends strongly on periodicity information, which is also a major cue for lexical tone recognition. As such, a sine-wave vocoder was used in the present CI simulations. For vocoded speech, the input acoustic signal was band-pass filtered into four frequency bands using fourth-order Butterworth filters. The cutoff frequencies of the analysis bands were 200, 591, 1426, 3205, and 7000 Hz, respectively. The amplitude envelope was extracted from each band by half-wave rectification and low-pass filtering (fourth-order Butterworth) with a 160 Hz cutoff frequency; previous studies have not shown any significant difference in speech understanding envelope filter frequencies above 128 Hz (e.g., Xu *et al.*, 2002), for a variety of perceptual tasks. The extracted envelope from each band was used to modulate sine-wave carriers whose center frequencies were the geometric center frequencies of the analysis bands. Finally, the modulated carriers were summed and normalized to have the same long-term root-mean-square as the input speech signal.

All CI listeners were tested using their everyday, clinical speech processors and settings. Subjects were instructed to use their everyday volume and microphone sensitivity settings, and to use the clinical map for everyday normal conversation instead of special maps for noise cancellation, music, distant speech, etc. Once set, subjects were asked to not change these settings during testing.

For both NH and CI subjects, stimuli were presented in a 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 the sentence as accurately as possible. The experimenter calculated the percent of words correctly identified in sentences. All words in the MEST materials were scored, resulting in a total of 70 key words for each list. No training or trial-by-trial feedback was provided during testing.

For NH listeners listening to four-channel simulation, the listeners listened to the four-channel vocoded speech for 30 min to minimize the effect of the procedural learning (familiarization) before formal test. For each condition, NH subjects were tested using two lists. Due to the time restrictions, the Mandarin-speaking CI subjects were tested using only one list per condition so all tests could be completed within a two hour test session. All experimental conditions (slow, fast, normal, and whispering) and testing sentence lists (lists 1–10) were randomized and counterbalanced across all subjects.

3. Results

NH subjects scored 100% correct with the original, unprocessed, experimental speech materials. Figure 1 shows individual data points and a box plot of sentence recognition scores for NH subjects listening to four-channel CI simulation (Panel A) and CI subjects listening to the original, unprocessed speech (Panel B).

For both NH listeners and CI subjects, overall performance was best for the slow-rate MEST sentences and gradually worsened as the speaking rate was increased. Performance sharply declined for whispered speech. One-way repeated-measures analyses of

Fig. 1. Sentence recognition scores as a function of different sentence materials. NH subjects were tested while listening to a four-channel CI simulation (A); CI subjects were tested with their clinical processors (B). The upper edge of the box indicates the 95th percentile of the data set, and the low edge indicates the 5th percentile. The solid lines in the box indicate the median recognition scores and the dashed lines indicate the mean recognition scores. The open symbols indicate the data points from individual subjects.

variance (RM ANOVAs) were performed for the NH simulation data and for the CI data, with test materials (fast, normal, slow, and whispered) as the treatment factor. There was a significant effect of test materials for both NH listeners [$F(3,31) = 24.59$, $p < 0.001$] and CI subjects [$F(3,51) = 37.33$, $p < 0.001$]. For NH listeners, post-hoc Holm–Sidak corrected pair-wise comparisons showed that recognition scores for whispered speech were significantly lower than for the other speech materials ($p < 0.001$), and that there was no significant difference between the remaining test materials. For CI subjects, post-hoc Holm–Sidak corrected pair-wise comparisons showed that recognition scores with whispered speech were significantly lower than with the other speech materials ($p < 0.001$), and that recognition scores with fast speech were significantly lower than with slow speech ($p < 0.001$). A two-way ANOVA was further performed on the data, with subject group (NH 4-channel simulation and CI) and test materials as treatment factors. There were significant effects for subject group [$F(1,83) = 4.85$, $p = 0.031$] and test materials [$F(3,83) = 16.35$, $p < 0.001$]; there was no significant interaction [$F(3,83) = 0.04$, $p = 0.989$].

Linear regression analyses showed strong correlations between the four speech measures (for all comparisons: $r > 0.82$, $p < 0.001$). Linear regression analysis also showed that duration of deafness was significantly correlated with performance with the fast-rate speech ($r = 0.61$, $p = 0.027$); duration of deafness was not correlated with the remaining speech measures. Age at testing was also not correlated with any of the speech measures. There were no clear patterns of results for the different etiologies of deafness or CI device type.

4. Discussion

The present results showed that Mandarin-speaking Chinese CI users' speech recognition was significantly affected by speaking style. While whispered speech had a strong effect of sentence recognition, speaking rate had relatively minor effect on performance. In the following we discuss the results in greater detail.

NH listeners were able to perfectly understand all of the unprocessed experimental speech materials, despite differences in speaking rate, the availability of F0 cues, etc. NH performance with the four-channel CI simulation was significantly better than that of CI subjects. Because of the limited spectral resolution, CI listeners can effectively access only 4–8 spectral channels (Shannon et al., 2004). While there were only four channels available in the CI simulation, near perfect recognition scores (ranged from 93% to 100%) were achieved in all NH listeners when listening to slow-

rate speech. However, the recognition scores of CI subjects were much more variable when compared to those of NH listeners. Near perfect recognition scores (range: 96%–100% correct) were observed in 8 out of 13 CI subjects. The remaining five CI subjects did not perform nearly as well, with scores ranging from 64% to 86% correct. Patient-related factors (e.g., poor nerve survival, shallow insertion depth, etc.) may have contributed to the poorer performance in these CI subjects. The poor spectral resolution in the CI simulation reduced NH subjects' speech understanding, compared to the perfect performance with unprocessed speech. In this sense, NH and CI subjects were similarly susceptible to reduced spectral resolution, resulting in similar performance patterns with the different sentence materials.

Recognition scores of whispered speech dropped by nearly 40 points, relative to performance with the comparable normal speaking rate. One possible explanation for this large drop is due to the missing periodicity cues in the whispered speech. With only amplitude envelope cues (2–50 Hz), lexical tone recognition has been shown to be approximately 60% correct for Mandarin-speaking NH subjects listening to whispered (Liang, 1963) or vocoded speech (Fu *et al.*, 1998). Lexical tone recognition has been shown to improve from 60% to 80% correct when temporal periodicity cues are available, significantly improving sentence recognition (Fu *et al.*, 1998). The poor performance with whispered speech in the present study further highlights the importance of F0 cues to Chinese tone and sentence recognition by Mandarin-speaking CI users.

Another interesting observation is that the performance deficit with whispered speech (40 points) was nearly twice as large as that reported by Fu *et al.* (1998) after periodicity cues were removed (i.e., 500 Hz vs. 50 Hz temporal envelope cutoff frequency), suggesting periodicity cues alone may not explain differences in speech understanding between normal and whispered speech. While whispering, voiceless sounds are produced as normal; voiced sounds (e.g., vowels) are produced by forcing air through a narrow glottal opening formed by holding the vocal cords rigid and close together. With sufficient spectral resolution (e.g., the original, unprocessed whispered speech), such acoustic distortions might be compensated by redundant speech cues. With reduced spectral resolution, such compensatory mechanisms may not be available. Ito *et al.* (2005) found that there is an upward shift in vowel formant frequencies for whispered speech, compared to normal speech. Voiced consonants in whispered speech have lower energy at low frequencies (below 1.5 kHz), with greater spectral flatness compared to normal speech. Again, CI subjects may be unable to compensate for such acoustic distortions due to the limited spectral resolution, at least for acutely measured performance. However, Ito *et al.* (2005) also found that training with a small sample of whispered speech utterances from a target talker significantly improved whispered speech recognition for that talker, suggesting that the poorer speech understanding with whispered speech may be overcome by training.

Although all CI subjects' performances dropped dramatically with whispered speech, the effect of speaking rate was much less consistent across CI subjects. There was large intersubject variability in terms of the effect of speaking rate. More than half of the CI users (7 out of 13 subjects) exhibited no degradation in performance with the normal or fast rate, relative to slow-rate speech. The remaining CI subjects were more strongly affected by speaking rate. Linear regression analysis showed that duration of deafness was significantly correlated with performance with fast-rate speech. There were no significant correlations between any of the other subject demographics and the other speaking styles. It is possible that the large deficits observed with fast-rate speech may be due to subjects' temporal processing deficits. Fu (2002) found a significant correlation between modulation detection and CI subjects' vowel and consonant recognition. Additional psychophysical measures may show a similar relationship between temporal processing and perception of fast-rate speech.

The present data also offer some insights regarding test materials for CI users. Most CI speech studies have measured performance with one or more types of clear-speech test materials (e.g., HINT or IEEE sentences, CNC words, phonemes, etc.).

Few studies have examined the effects of speaking style within one or more speech tests (e.g., clear versus conversational, speech rate, whispered versus normal). For the same speech materials, different recordings have been used for speech perception tests in various studies without any description about the speaking style for the recordings. It is possible that speech recordings may greatly differ in terms of speaking style, especially speaking rate. For example, the speaking rate (4.46 words/s) of the recently developed Mandarin Hearing in Noise Test (MHINT, Wong *et al.*, 2007) was between the normal (3.64 words per second) and fast rate (5.67 words per second) of MEST sentences. The data from the present study suggested that speaking styles, such as speaking rate, should be carefully considered and described in detail when evaluating CI patients' speech recognition performances, or when comparing data from different studies that use different recordings of the same test materials.

Acknowledgments

The authors thank all the subjects who participated in this study. The authors also thank John J. Galvin III for editorial assistance. This work was partially supported by NIH Grant No. DC004993.

References and links

- Chang, Y.-P., and Fu, Q.-J. (2006) "Effects of talker variability on vowel recognition in cochlear implants," *J. Speech Lang. Hear. Res.* **49**, 1331–1341.
- Eskenazi, M. (1993). "Trends in speaking styles research," in *EUROSPEECH'93*, pp. 501–509.
- Fu, Q.-J. (2002). Temporal processing and speech recognition in cochlear implant users, *Neuroreport* **13**, 1635–1640.
- 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.
- Fu, Q.-J., Chinchilla, S. Nogaki, G., and Galvin, J.J., III (2005). "Voice gender discrimination: the role of periodicity and spectral profile," *J. Acoust. Soc. Am.* **118**, 1711–1718.
- Ito, T. Takeda, K., and Itakura, F. (2005) "Analysis and recognition of whispered speech," *Speech Commun.* **45**, 139–152.
- Liang, Z.-A. (1963). "The auditory perception of Mandarin Tones," *Acta Phys. Sin.* **26**, 85–91.
- Lin, M.-C. (1988). "The acoustic characteristics and perceptual cues of tones in Standard Chinese," *Chin. Yuwen* **204**, 182–193.
- Liu, C. Galvin, J.J. 3rd, Fu, Q.-J., and Narayanan, S.S. (2008). "Effect of spectral normalization on different talker speech recognition by cochlear implant users," *J. Acoust. Soc. Am.* **123**, 2836–2847.
- Liu, S. Del Rio, E. Bradlow, A. R., and Zeng, F.G. (2004). "Clear speech perception in acoustic and electric hearing," *J. Acoust. Soc. Am.* **116**, 2374–2383.
- Picheny, M. A. Durlach, N. I., and Braida, L.D. (1985). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *J. Speech Hear. Res.* **28**, 96–103.
- Shannon, R. V. Fu, Q. J., and Galvin J, J. III (2004). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," *Acta Otolaryngol. Suppl.* **552**, 50–54.
- Sommers, M. Nygaard, L., and Pisoni, D. (1992). "Stimulus variability and the perception of spoken words: Effects of variation in speaking rate and overall amplitude," in *ICSLP '92*, pp. 217–220.
- Uchanski, R. M. Choi, S. S. Braida, L. D. Reed, C. M., and Durlach, N.I. (1996). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," *J. Speech Hear. Res.* **39**, 494–509.
- Wong, L. L. Soli, S. D. Liu, S. Han, N., and Huang, M.W. (2007). "Development of the Mandarin Hearing in Noise Test (MHINT)," *Ear Hear.* **28** (2 Suppl.), 70S–74S.
- Woodland, P.C. (2001). "Speaker Adaptation for Continuous Density HMMs: A Review," in *ITRW on Adaptation Methods for Speech Recognition*, pp. 11–19.
- Xu, L. Tsai, Y., and Pfingst, B. (2002). "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *J. Acoust. Soc. Am.* **112**, 247–258.