Evaluation and analysis of whispered speech for cochlear implant users: Gender identification and intelligibility

Oldooz Hazrati,[1,a)] Hussnain Ali,[1] John H. L. Hansen,[1] and Emily Tobey[2]
[1]*Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas 75080, USA*
[2]*School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA*

(Received 11 August 2014; revised 27 February 2015; accepted 19 May 2015; published online 2 July 2015)

This study investigates the degree to which whispered speech impacts speech perception and gender identification in cochlear implant (CI) users. Listening experiments with six CI subjects under neutral and whispered speech conditions using sentences from the UT-Vocal Effort II corpus (recordings from male and female speakers) were conducted. Results indicated a significant effect of whispering on gender identification and speech intelligibility scores. In addition, no significant effect of talker gender on the speech/gender identification scores was observed. Results also suggested that exposure to longer speech stimuli, and consequently more temporal cues, would not improve gender identification performance in CI subjects. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4922230]

[MSS] Pages: 74–79

I. INTRODUCTION

Cochlear implant (CI) devices, in general, enable profoundly deafened individuals to identify uncorrupted and neutral speech to a great extent. CI users understand spectrally reduced speech (SRS), speech processed through CIs, quite well as long as no distortion is introduced which compromises the speech signal (Dorman *et al.*, 2000). Distortions may be associated with the listening environment due to room reverberation and background noise (Hazrati *et al.*, 2013; Hazrati, 2012), or due to the speaking style of speakers, for example, shouted or whispered speech (Luo *et al.*, 2006). The ability of CI users in recognizing speech drops substantially in challenging listening scenarios due to the limited temporal and spectral resolution of speech signal processed in the CI devices (Dorman *et al.*, 1998; Loizou *et al.*, 2009). Whispered speech mode presents additional unique challenges to CI users with electric hearing.

Whisper-based speech mode is fundamentally different in speech production versus neutral speech. Whispered speech lacks pitch and voicing cues due to the stationarity of vocal folds while whispering. Consequently, periodic pulsing information, which in turn results in pitch perception of whispered speech, is not present, therefore the perceived pitch is highly correlated with the frequency of the first two formants (especially F_2) (Thomas, 1969). In normal-hearing (NH) listeners, despite the lack of such periodic pulsing structure, there still remains some degree of pitch perception resulting in reasonable gender identification and speech recognition. However, speech perception in CI recipients primarily relies on the spectrally reduced speech presented through their devices. Hence, even subtle variations due to speech prosody could result in significant changes in CI listeners' performance.

Whispering, in particular, changes the temporal fine structure of speech but does not degrade spectral details representing the vocal tract response (Freyman *et al.*, 2012). Bosker *et al.* (2010) hypothesized that due to the high contribution of consonants to speech intelligibility and the fact that intensity of vowels and consonants are equal in whispered speech, whispering may improve speech intelligibility in CIs. However, their hypothesis was neither proved nor disproved after testing NH listeners with vocoded sentences spoken in Dutch. Freyman *et al.* (2012) studied the intelligibility of whispered speech in stationary and modulated noise conditions with NH listeners. They found that the intelligibility of whispered speech was significantly lower than that of the neutral speech in both stationary and modulated noise types. Moreover, they observed that the masking release was reduced when comparing vocoded speech to whispered speech (Freyman *et al.*, 2012). Listening experiments with NH subjects indicated that the perceived pitch of vowels closely corresponded to the formant frequencies (F_1 and F_2) (Thomas, 1969; Higashikawa *et al.*, 1996). Despite the assumption made in (Bosker *et al.*, 2010), taking into account the spectral and periodic temporal fine structure degradation of vocoded speech as well as perceived pitch mismatches caused by whispering, it is expected that whispering significantly drops the intelligibility of speech for CI listeners.

In addition to speech understanding, speaker and/or gender identification is an important factor in recovering linguistic information. It has been shown that fundamental and formant frequencies (f_0, F_1, and F_2) are important factors in speaker-gender classification (Bachorowski and Owren, 1999). Due to the reduced temporal and spectral cues of CI processed speech, gender identification performance of CI users drop even under quiet and neutral speech conditions (Fu *et al.*, 2005). As noted earlier, the perceived pitch of whispered vowels correspond to the first and second formant frequencies (Thomas, 1969); therefore, gender identification

[a)]Electronic mail: hazrati@utdallas.edu

 0001-4966/2015/138(1)/74/6/$30.00

in whispered speech becomes challenging for CI listeners with SRS. The voice gender identification was studied in (Fu *et al.*, 2005) with CI and NH subjects for the monophthongs and diphthongs in a "/*h*/-*vowel*-/*d*/" context. It was concluded that CI subjects may rely more on the temporal cues for voice gender identification task due to their SRS perception.

Based on the above noted findings and the hypothesis made in (Bosker *et al.*, 2010) regarding the intelligibility benefit provided by whispering for CI listeners, the present study investigated speaker-gender identification, and speech intelligibility for CI users presented with sentences in neutral and whispered speech conditions. In addition, our goal is to address the following main question: If CI listeners rely mostly on the temporal cues (due to SRS), will access to more cues regarding different phonemes of a sentence result in a whispered gender identification close to their neutral scores, or does the diversity of temporal cues provide insufficient information/knowledge in order to compensate for the lack of periodicity cues?

In order to further analyze the combined effects of whispering and spectral distortion of the CI processed speech on both speech and speaker-gender identification, interactions between the subjects' MAP (stimulation rate, electric dynamic range) and their test scores were evaluated.[1]

II. METHODS

A. Subjects

Six post-lingually deafened CI users (three males and three females), with an average age of 62.8 yr, were recruited for this study. All participants were native speakers of American English and had a minimum of three-year experience with their devices. All subjects used Cochlear manufactured devices with Nucleus 5 processors and had at least 20 active electrodes. Participants used the Advanced Combination Encoder (ACE) speech coding strategy, with an eight maxima selection procedure at each analysis cycle, on a daily basis. Details on CI users demographics and clinical MAP are provided in Table I.

B. Speech material

Speech sentences from the UT-Vocal Effort II (VEII) corpus (Zhang and Hansen, 2009) were used for both gender identification and intelligibility listening tests. The UT-VEII corpus comprises "read" and "spontaneous" speech from 112 speakers (37 male and 75 female). In the read portion of UT-VEII, each subject speaks 41 sentence prompts from the TIMIT database (Zue *et al.*, 1990) and two newspaper paragraphs while alternating between neutral and whispered mode. The recordings were conducted in an ASHA-certified single-walled sound booth. Speech signal was captured using a head-worn close talking microphone Shure Beta 53 and recorded using a Fostex 8 D824 digital recorder at 44.1 kHz/16 bits per sample and downsampled to 16 kHz for this study. The root-mean-square energy of all sentences was equalized to the same value corresponding to approximately 65 dBA.

C. Procedure

Subjects were tested using the Personal Digital Assistant (PDA)-based cochlear implant research platform (Ali *et al.*, 2013) in a double-walled sound proof booth. Recorded speech sentences from the UT-VEII database were processed in MATLAB and streamed to the subjects' implant via PDA interface. The processor was programmed with the ACE coding strategy and configured for the subject's threshold and comfortable levels, and stimulation parameters. Bilateral CI subjects were tested with the best ear (the ear participant reported to have more clear and intelligible speech). Training sentences were played prior to the test and volume adjustments were made to each individual's desired level. The presentation order of the sentences was randomized across subjects.

In this study, 20 phonetically balanced sentences (10 from a female and 10 from a male speaker) per condition were used for intelligibility listening tests. Similarly, 10 female and 10 male speakers (one sentence per speaker) were used as gender identification material in each condition (neutral/whispered). Due to the limited number of the read sentences in UT-VEII corpus, speech material presented in the intelligibility tests were reused, though randomized, in the gender identification task. Speech intelligibility experiments were conducted first for all CI users and followed by gender identification testing. In this manner, speech intelligibility scores were not affected by possible familiarity of listeners with sentences they previously heard.

For the gender identification task, the average fundamental frequency of male and female speakers was 118 and 176 Hz, respectively. The range of fundamental frequency of the speakers considered in this study, for both neutral and whispered conditions, is provided in Table II. A MATLAB implementation of the robust algorithm for pitch tracking (RAPT) is used for fundamental frequency estimation[2] (Talkin, 1995).

TABLE I. Demographics and clinical MAP parameters of the CI users tested.

Subject ID	Gender	Age (years)	Implant type	No. of active electrodes	Stimulation rate (Hz)	Average electric dynamic range	Years implanted
S1	M	60	CI512	21	900	38	3.5
S2	M	80	Freedom	21	1200	17	8
S3	F	55	Freedom	22	900	42	4.8
S4	F	56	CI512	20	900	40	3.8
S5	M	66	Freedom	21	900	35	4.2
S6	F	60	Freedom	20	500	10	3

TABLE II. Fundamental frequency (in Hz) of the UT-VEII corpus speakers used in this study.

	Whisper		Neutral	
Experiment	M	F	M	F
Gender ID	96–156	153–214	97–136	149–204
Intelligibility	114	198	117	197

III. EXPERIMENTS

A. Intelligibility listening test

For the intelligibility test, 20 sentences with an average duration of 4.8 s were each presented to the subjects twice and they were asked to repeat as many words as they could perceive. The number of words correctly identified divided by the total number of words was considered as the intelligibility score. Whispered speech was presented to three CI listeners as the first, and for the other three participants, as the last condition.

B. Gender identification test

For the gender identification test in each condition, 20 sentences spoken by different speakers (10 males and 10 females) were presented to CI participants. The listeners were instructed to identify the gender of the speaker, without concentrating on word recognition, when each sentence was played. No repetitions were allowed for this test. The number of speaker genders correctly identified was divided by the total number of speakers per condition to compute the speaker-gender identification score. Similar to the intelligibility listening test, half of the listeners were presented with neutral speech first and the other half with whispered speech first.

IV. RESULTS

Individual, as well as overall mean speech intelligibility results for neutral and whispered speech conditions are shown in Fig. 1. Speech intelligibility scores dropped from an average of 68.8% correct in the neutral condition to an average of 43.4% in the whispered condition. The speech identification of CI listeners decreased from an average of 67.2% and 70.4% in the neutral speech condition to an average of 39.3% and 47.4% in whispered speech condition when presented with speech sentences spoken by a female talker [Fig. 1(a)] and a male talker [Fig. 1(b)], respectively.

Repeated-measures analysis of variance (ANOVA) with α parameter set to 0.05 was performed to assess the effect of whispering on speech intelligibility scores of CI listeners. Subjects were considered as a blocked factor while intelligibility scores from male and female talkers in neutral and whispered conditions were used as the main analysis factors. ANOVA revealed a statistically significant effect for whispering ($F_{1,5} = 32.627$, $p = 0.002$) on the speech intelligibility scores compared to neutral speech. However, no significant effect of speaker-gender on

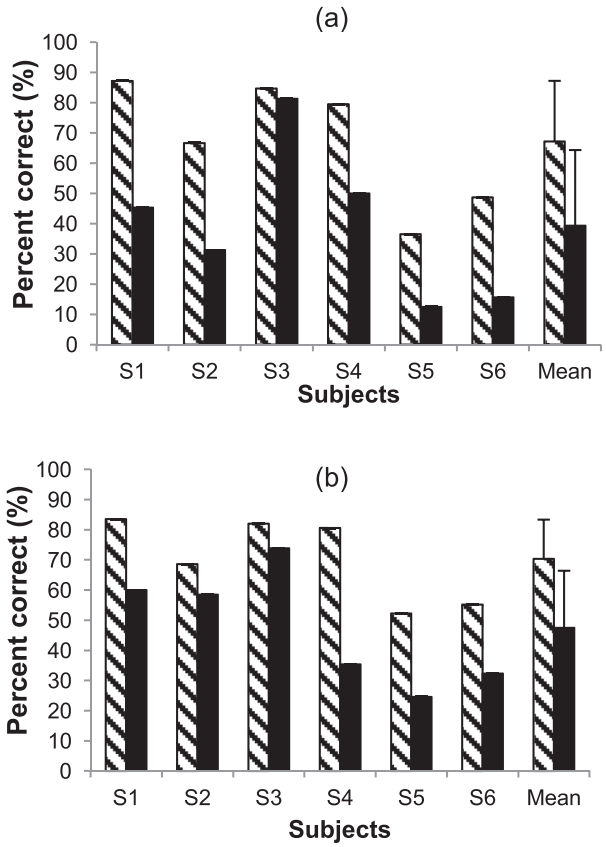

FIG. 1. (a) Female and (b) male, individual speech intelligibility scores in neutral (dashed bars) and whispered (solid bars) conditions. Also shown, mean and standard deviations.

intelligibility scores was observed ($F_{1,5} = 2.130$, $p = 0.204$). The interaction between speech type (neutral/whispered) and speaker gender was not significant ($F_{1,5} = 0.598$, $p = 0.472$). These results are in-line with our assumption on reduced intelligibility due to whispering.

Individual and average results for the speaker-gender identification task under neutral and whispered conditions are presented in Fig. 2 [female in 2(a), male in 2(b)]. The results indicate that CI users were able to correctly recognize speaker-gender 95% of the times when presented with neutral spoken sentences. However, their gender identification performance decreased to an average of 75% for whispered speech. Gender identification scores dropped from an average of 98.3% and 91.7% in neutral speech condition to an average of 83.8% and 66.7% in whispered speech condition for female [Fig. 2(a)] and male [Fig. 2(b)] talkers, respectively.

Repeated-measures ANOVA was also performed by considering speech type (neutral/whispered) and the speaker gender (male/female) as the analysis factors. Results indicated a significant effect of speech type on correct gender identification ($F_{1,5} = 34.286$, $p = 0.002$). However, no particular preference for correct male or female speaker identification was found ($F_{1,5} = 0.682$, $p = 0.447$). Also, no significant interaction between speech type and speaker gender was found ($F_{1,5} = 0.682$, $p = 0.447$). This demonstrates that

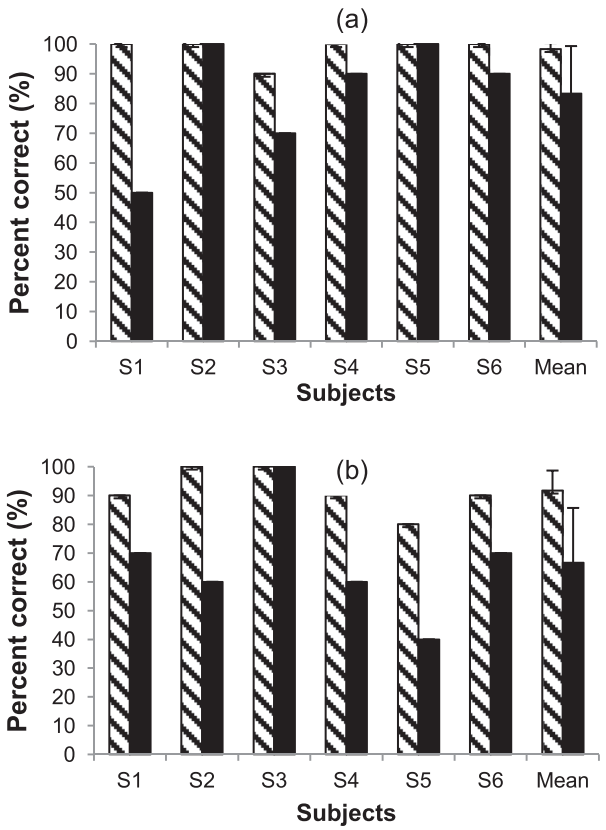

FIG. 2. (a) Female and (b) male, individual speaker-gender identification scores in neutral (dashed bars) and whispered (solid bars) conditions. Also shown, mean and standard deviations.

speech type (neutral/whispered) significantly influences the correct gender identification in CIs.

V. SUMMARY AND DISCUSSION

The study by Ringeling (1984) with NH listeners reported increased speech intelligibility with Dutch isolated words in whispered speech compared to neutral speech. However, Bosker *et al.* (2010) were unable to prove or disprove the hypothesis that whispering may be beneficial to the intelligibility of speech because of the increased intensity of consonants, compared to vowels, in whispered speech (NH listeners were tested with vocoded Dutch speech sentences). The results from the current study indicates otherwise, speech intelligibility on average dropped by 25% in the whispered condition for CI participants. Although inter-subject variability in intelligibility scores was observed (6% to 33% difference), reduced intelligibility in whispered condition was seen for all six participants. This contradiction in results could be attributed to the test protocol and linguistic content, as former studies (Bosker *et al.*, 2010; Ringeling, 1984) were conducted with NH listeners in Dutch language. It is possible that the spectrally reduced speech produced by CI devices may indeed emphasize the reduced periodicity cues effect caused by whispering and result in significant degradation of intelligibility of the speech for CI users.

The effect of whispering on the periodic pulsing and voicing cues is shown in Fig. 3. Spectrogram comparisons of the same TIMIT sentence spoken in neutral (top panel) and

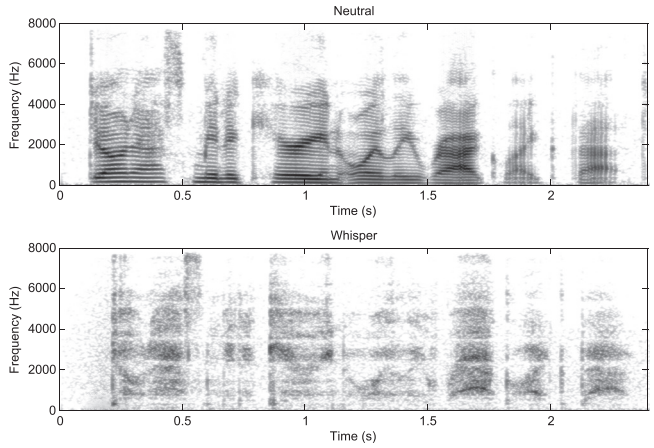

FIG. 3. Spectrograms of TIMIT sentence "*Do not ask me to carry an oily rag like that*" spoken in neutral (top) and whispered (bottom) style.

whisper (bottom panel) style indicates loss of pitch information, changes in spectral contrast and formants distribution in whispering (Ghaffarzadegan *et al.*, 2014). These effects are in turn translated to the electric stimuli generated by the CI processor. Figure 4 shows electrodograms of the same

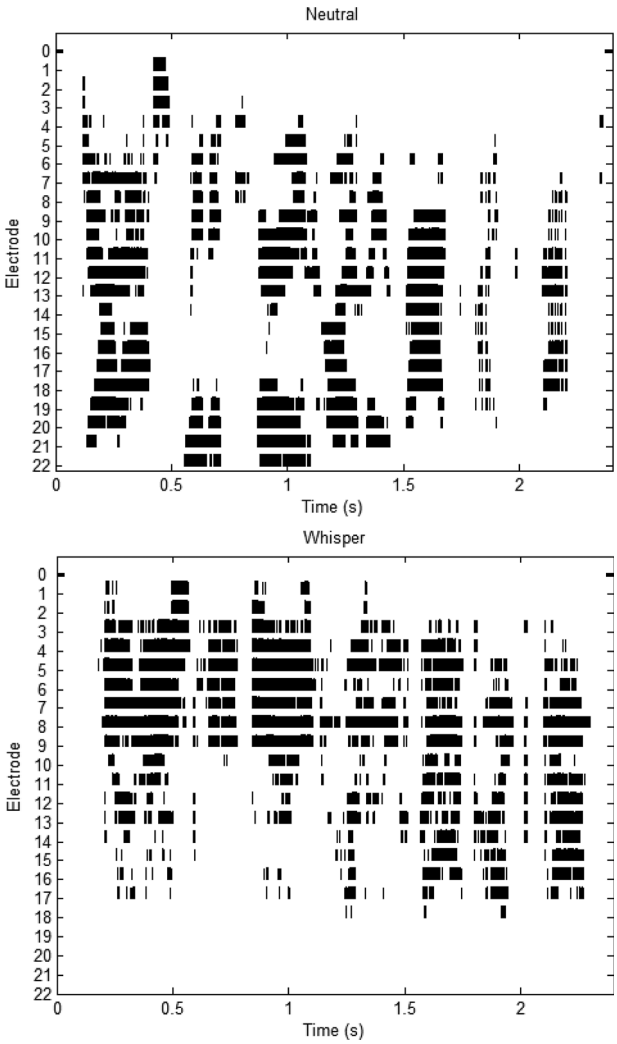

FIG. 4. Electrodograms of TIMIT sentence "*Do not ask me to carry an oily rag like that*" spoken in neutral (top) and whispered (bottom) style.

TIMIT sentence under neutral and whisper speaking style. Since ACE channel selection is based on n-of-m rule (*n* out of *m* channels with maximum energy are selected in each stimulation cycle), clearly high-frequency channels are favored in whispered mode due to increased consonants energy [Fig. 4(b)]. Therefore electrodes corresponding to frequencies up to 800 Hz (electrodes 18–22) which is the frequency band carrying pitch and first formant information, known to be critical for speech understanding, are never selected in whispered speaking style.

Research on voice perception has shown that pitch and spectral properties of speech provide strong cues to recognize gender and identity of the speaker (Gonzalez and Oliver, 2005). Therefore, as mentioned above, any distortion affecting pitch perception and spectral properties (such as whispering and CI processing of speech) will decrease speaker-gender identification performance (Massida *et al.*, 2013). As expected, the speaker-gender identification task was significantly influenced by speech type. Whispered speech had a significant negative impact on percentage correct responses for gender discrimination as compared to neutral speech. However, on average the scores were significantly above chance, which suggests the potential availability of some pitch cues in whispered speech.

In the current study, speech sentences instead of isolated words were used in the gender identification experiment in order to provide sufficient temporal context information to CI users regarding various phonemes to determine whether having access to longer duration of speech will compensate for the lack of periodicity cues, and if this will result in a nonsignificant speaker-gender identification difference between neutral and whispered speech. Speaker-gender identification results suggest that the negative impact of reduced periodicity cues on speaker gender recognition could not be completely suppressed even when presented with phonetically rich sentences.

Our experimental results indicate different patterns between speech intelligibility and gender identification scores under neutral and whispered speech conditions (implying that one subject may do better in gender identification under whisper condition but may perform poor in speech understanding in the same condition). It may be that, due to the different nature of speech intelligibility and gender identification tasks, where the former relies mostly on consonant recognition and the latter relies on periodicity cues which are mostly available in vowels, each CI user may demonstrate speech intelligibility and speaker-gender recognition results in-line or in-contrast to each other. These patterns may be attributed to the differences in electrode insertion depth, number of surviving neurons in the cochlea, or other top-down mechanisms among cochlear implantees.

No evident trend/pattern between speech intelligibility or gender identification results and the dynamic range variability among CI users was observed (for both neutral and whispered speech). Note that, due to the limited number of CI participants and subjective variability, it was not possible to make valid judgments on correlations between subject's intelligibility/gender identification scores and their processor type, stimulation parameters, and electric dynamic range.

VI. CONCLUSIONS

The present study evaluated the effect of whispering on speech intelligibility and speaker-gender identification performance in CI users. Six cochlear implantees were presented with sentences from the UT-Vocal Effort II corpus which contained recordings from both male and female speakers under neutral and whispered conditions. Both speech recognition and gender identification scores were substantially reduced in whispered speech. Experimental results indicated a significant effect of whisper on both speech intelligibility and gender identification performance. In addition, no significant effect of talker gender on the speech/gender identification scores was observed. The results also suggested that exposure to longer speech stimuli and consequently more temporal cues may not completely compensate for degraded pitch and voicing cues effect on the gender identification performance in CI subjects. It is possible that different patterns of speech intelligibility and speaker-gender identification scores were due to the differences in insertion depth of electrode array (place) and surviving neurons among CI users. Therefore, a good/poor performance in either task (speech and gender recognition) may not necessarily guarantee equivalent performance in the other.

ACKNOWLEDGMENTS

This work was supported by grant R01 DC010494 awarded from NIH (PI: E.T.). The authors would like to thank the CI users for their time.

[1]Part of this study was presented at CI 2014 Conference, 18–21 June, Munich, Germany.

[2]It is worth mentioning that the reported *estimated* F0s for different speakers were computed only using each speaker's neutral speech.

Ali, H., Lobo, A. P., and Loizou, P. C. (**2013**). "Design and evaluation of a PDA-based research platform for cochlear implants," IEEE Trans. Biomed. Eng. **60**(11), 3060–3073.

Bachorowski, J. A., and Owren, M. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106**(2), 1054–1063.

Bosker, H. R., Briaire, J., Heeren, W., Heuven, V. J., and Jongman, S. (**2010**). "Whispered speech as input for cochlear implants," Linguistics Netherlands **27**(1), 1–15.

Dorman, M., Loizou, P. C., and Fitzke, J. (**1998**). "The identification speech in noise by cochlear implant patients and normal-hearing listeners using 6-channel signal processors," Ear Hear. **19**(6), 481–484.

Dorman, M., Loizou, P., Fitzke, J., and Tu, Z. (**2000**). "Recognition of monosyllabic words by cochlear implant patients and by normal-hearing subjects listening to words processed through cochlear implant signal processing strategies," Annal. Oto., Rhin., Laryng. **109**(12), Suppl. 185, 64–66.

Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (**2012**). "Intelligibility of whispered speech in stationary and modulated noise maskers," J. Acoust. Soc. Am. **132**(4), 2514–2523.

Fu, Q. J., Chinchilla, S., Nogaki, G., and Galvin, J. J. III (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**(3), 1711–1718.

Ghaffarzadegan, S., Boril, H., and Hansen, J. H. L. (**2014**). "Model and feature based compensation for whispered speech recognition," in *Proc. INTERSPEECH*, pp. 2420–2424.

Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**(1), 461–470.

Hazrati, O. (**2012**). "Development of dereverberation algorithms for improved speech intelligibility by cochlear implant users," Ph.D. dissertation, University of Texas at Dallas.
Hazrati, O., Sadjadi, S. O., Loizou, P. C., and Hansen, J. H. L. (**2013**). "Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy," J. Acoust. Soc. Am. **134**(5), 3759–3765.
Higashikawa, M., Nakai, K., Sakakura, A., and Takahashi, H. (**1996**). "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study for the Fourier transform," J. Voice **10**(2), 155–158.
Loizou, P. C., Hu, Y., Litovsky, P., Yu, G., Peters, R., Lake, J., and Roland, P. (**2009**). "Speech recognition by bilateral cochlear implant users in a cocktail party setting," J. Acoust. Soc. Am. **125**(1), 372–383.
Luo, X., Fu, Q-J., and Galvin, J. J. (**2006**). "Vocal emotion recognition with cochlear implants," in *Proc. INTERSPEECH*, pp. 1830–1833.
Massida, Z., Marx, M., Belin, P., James, C., Fraysse, B., Barone, P., and Deguine, O. (**2013**). "Gender categorization in cochlear implant users," J. Speech, Lang., Hear. Res. **56**(5), 1389–1401.
Ringeling, J. C. T. (**1984**). "Reducing redundancy in normal, soft and whispered speech: A study on native and near-native perception," Ph.D. dissertation, Utrecht University.
Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier, New York), pp. 495–518.
Thomas, I. B. (**1969**). "Perceived pitch of whispered vowels," J. Acoust. Soc. Am. **46**(2), 468–470.
Zhang, C., and Hansen, J. H. L. (**2009**). "Advancement in whisper-island detection with normally phonated audio streams," in *Proc. INTERSPEECH*, pp. 860–863.
Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**(4), 351–356.