Running title: Talkers’ voice quality through cochlear implant

Simulated cochlear-implant processing reveals major loss of acoustic information relevant to talkers’ voice quality differences

Meisam K. Arjmandi $^{1,b}$, Hamzeh Ghasemzadeh $^{1,c}$, Laura C. Dilley $^{1}$

$^{1}$ Department of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Road, East Lansing, Michigan 48824, USA

Date of manuscript submission: June 3, 2020

$^{a)}$ Portions of this work were presented at the 177th Meeting of the Acoustical Society of America, Louisville, Kentucky, USA, May 2019.

$^{b)}$ Current affiliation: Department of Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear and Harvard Medical School, Boston, Massachusetts, USA. Email: meisam_khalilarjmandi@meei.harvard.edu

$^{c)}$ Also affiliated with: The Department of Computational Mathematics Science and Engineering, Michigan State University, East Lansing, Michigan, USA.
ABSTRACT

The ability to discern variations in talkers’ voice quality is important for effective talker identification and robust speech processing; yet, little is known about how faithfully acoustic information relevant to variations in talkers’ voice quality is transmitted through cochlear implant (CI) speech processing. This study analyzed unprocessed and CI-simulated versions of sustained vowel sounds /a/ from two groups of individuals with normal and disordered voice qualities to investigate the effects of CI speech processing on acoustic information relevant to the talkers’ voice quality distinction. The CI-simulated stimuli were created by processing the vowel sounds using 4-, 8-, 12-, 16-, 22-, and 32-channel noise-vocoders. The voice quality for each stimulus was characterized by calculating *mel-frequency cepstral coefficients* (MFCCs). Then, the effects of CI speech processing on the acoustic distinctiveness between normal and disordered voices was measured by calculating the *Mahalanobis distance* and classification accuracy of *support vector machines* (SVMs) on their MFCC features. The results showed that CI noise vocoding is substantially detrimental to acoustic information involved in voice quality distinction, suggesting that CI listeners likely experience difficulties in perceiving voice quality variations. The results underscore challenges CI users may face for effective recognition of talkers and processing their speech.

*Keywords*: Talkers’ voice quality, cochlear implant, acoustic distance
I. INTRODUCTION

The abundant spectro-temporal cues in speech serve as rich sources of information for listeners to learn and retrieve a variety of linguistic and indexical cues important for robust speech comprehension and language development. Talkers’ voice quality is one such aspect that listeners with normal hearing (NH) have access for capturing several indexical and sociolinguistics information and, thus establish successful spoken communication. Variations in talkers’ voice quality facilitate human spoken communication in multiple ways. These variations may provide perceptually salient grammatical and phonological cues for language comprehension (Cameron, 2001; Dicanio, 2009; Dolar, 2006; Garellek & Keating, 2011; Gordon, 2001; Gordon & Ladefoged, 2001; Henton, 1986; Ogden, 2001). Furthermore, listeners use acoustic information relevant to talkers’ voice quality to encode and discern various indexical information related to talkers’ identity such as gender, age, and race (Abberton & Fourcin, 1978; Eidsheim, 2012; Laver, 1968). This encoded information is actively incorporated by listeners to identify and recognize talkers (Gussenhoven & Rietveld, 1998; Latinus & Belin, 2011; Van Dommelen, 1987) and understand their speech (e.g., Creel & Tumlin, 2011). Talkers’ voice quality also contributes to speech understanding through constructing stance in communicative interaction, particularly by conveying talkers’ affective states during communication (e.g., happiness, anger, etc.; Aubergé & Cathiard, 2003; Guzman, Correa, Muñoz, & Mayerhoff, 2013; Podesva, 2007; Sicoli, 2010; Tsai et al., 2010; Zimman, 2012). In addition, listeners use talkers’ voice quality to infer their medical conditions (i.e., physical, psychological, and mental status; e.g., Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Laver, 1968). While listeners with normal hearing (NH) can often readily infer this information from talkers’ voice, little is known about how CI speech processing may impact faithful transmission of acoustic information relevant to talkers’ voice quality distinction.
The present study examined unprocessed and CI-simulated sustained vowel sound /a/ uttered by two groups of individuals with normal and disordered voice qualities to examine the effect of CI speech processing on voice quality distinction.

A. Talkers’ voice quality and robust speech comprehension

Acoustic information relevant to talkers’ voice quality signals a range of attributes, which together facilitates robust perception of speech. In facts, variations in talkers’ voice translate into perceptual attributes of voice quality such as breathiness, roughness, creakiness, and nasality (Childers & Lee, 1991; Eskenazi, Childers, & Hicks, 1990; Klatt & Klatt, 1990), which can directly provide linguistic cues for speech sound contrast (Dicanio, 2009; Garellek & Keating, 2011; Gordon, 2001; Gordon & Ladefoged, 2001). For instance, speakers of English and Finish modify their voicing behaviors and use creaky voice to signal phrase-final position (Henton, 1986; Ogden, 2001) and convey linguistic information at segmental and prosodic levels (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Dilley, Arjmandi, Ireland, Heffner, & Pitt, 2016; Redi & Shattuck-Hufnagel, 2001). Spectral measures such as $H1-H2$, $H1-A2$, and cepstral peak prominence (CPP) are acoustic correlates of talkers’ voice quality that were shown to be able to reflect these variations and distinguish between phonation types (Dicanio, 2009; Garellek & Keating, 2011). Furthermore, access to acoustic information related to talkers’ voice quality is crucial for effective speech comprehension because learning talker-specific information is essential for dealing with talker variability and, thus robust perception of speech (Johnson, 2005; Kleinschmidt & Jaeger, 2015; Pisoni, 1992). These findings highlight the importance of having access to voice quality-related acoustic information in talkers’ voice for robust speech comprehension.
B. Talkers’ voice quality and talker identification

Robust speech comprehension is also indirectly facilitated through talker identification and recognition. Listeners actively attend to variations in the talkers’ voice quality to index variability across talkers in their identity. Several studies have demonstrated the connection between talkers’ voice quality and their gender (e.g., Gussenhoven, 2004; Ohala, 1983; Puts, Hodges, Cárdenas, & Gaulin, 2007), race (Alim, 2004; Irwin, 1977; Moisik, 2013; Thomas & Reaser, 2004), ethnicity (Purnell, Idsardi, & Baugh, 1999), social group (Esling, 1978; Sicoli, 2007; Stuart-Smith, 1999), and cultural status (Rilliard, Shochi, Martin, Erickson, & Aubergé, 2009; Stross, 2013), which are all indices of talkers’ identity (Eidsheim, 2012; Podesva & Callier, 2015). Perceptual and acoustic correlates of voice quality such as pitch, phonation (e.g., breathiness, harshness, and creakiness), formant frequencies, and difference between amplitude of the first and second harmonics ($H1-H2$) have been shown to reflect talkers’ anatomical differences in the size of their larynx and vocal tracts, and thus are incorporated by listeners to identify talkers’ voice (Henton & Bladon, 1985; Ohala, 1983; Simpson, 2009). Therefore, preserving the integrity of spectro-temporal information in talkers’ voice is important for robust understanding of their speech by listeners.

Listeners are able to determine talkers’ gender from listening to isolated vowels (Coleman, 1971), as well as to continuous speech (Gelfer & Bennett, 2013). In addition, studies on African American English demonstrated a connection between talkers’ voice quality and their race (Irwin, 1977; Thomas & Reaser, 2004). For example, non-modal voice qualities were found to be frequently used by African American talkers, leading to a relatively harsh voice quality (Alim, 2004; Britt, 2011). Purnell et al., (1999) showed that landlords use acoustic information from a short segment of speech in the word “hello” to identify tenants with African American racial background and inappropriately discriminate against them. Acoustic analysis of African American
vowels showed that this racial judgment is likely facilitated through acoustic features relevant to talkers’ phonatory and articulatory behaviors (Arjmandi, Dilley, & Wagner, 2018; Thomas, 2007). Recognition of talkers’ social class is another dimension whereby voice quality contributes to the recognition of talkers’ identity (Esling, 1978; Henton, 1986; Stuart-Smith, 1999). For example, the level of harshness and pharyngealization in talkers’ voice were found to be associated with talkers’ social class. It is therefore important for listeners to have access to acoustic information relevant to talkers’ voice quality variations for talker identification and recognition, and thus effective speech understanding.

C. The role of talkers’ voice quality in language development

Learning talker’s voice quality also plays an important role in spoken language development. Children and adults incorporate acoustic information relevant to talkers’ voice quality to develop and construct talker-specific identity. The ability to map talker-specific acoustic information to talker identity is fundamental to children’s successful language development. Infants use talker-specific information for word recognition and language learning (Houston & Jusczyk, 2000, 2003; Ramírez-Esparza, García-Sierra, & Kuhl, 2017), suggesting that they likely develop mental representations specific to talkers’ voice information. Time-course of attending to talker-specific cues in speech begins very early before birth where infants recognize their mother’s voice (Kisilevsky et al., 2003) and was also shown in newborn infants who preferred listening to their mothers’ voice over that of a stranger (DeCasper & Fifer, 1980). Models of long-term memory for voice quality perception suggest that listeners encode and use patterns of voice quality variations in talkers’ speech for perception of voice quality (Papcun, Kreiman, & Davis, 2005). Talkers’ voice quality contributes to learning several other information about talkers such as their physical status (e.g., body size; Podesva & Callier, 2015), behavioral traits (e.g., smoking behaviors), and
their physical, psychological and mental health (e.g., vocal folds disorders, anxiety level, mood; e.g., Kreiman, Vanlancker-Sidtis, & Gerratt, 2005). One example of relying on voice quality for medical diagnosis is the auditory perceptual assessment of individuals’ voice by speech-language pathologists (SLPs) for differential diagnosis of disordered voices (Kreiman & Gerratt, 1996; Kreiman, Gerratt, & Ito, 2007). Although some of this information is partially retrievable in the absence of voice quality cues (e.g., Fellowes, Remez, & Rubin, 1997), the significant contribution of these cues in multiple aspects from establishing robust and effective spoken communications to differential diagnosis of individuals with disordered voice quality is indisputable.

**D. Talkers’ voice quality through CIs**

While listeners with NH have access to acoustic information related to talkers’ voice quality, it is yet unknown how users of cochlear implants may be at disadvantage in receiving this information due to limited spectral resolution in CI speech processing. Cochlear implants device has enabled many individuals with severe-to-profound sensorineural hearing loss to experience a partially restored hearing and develop language. However, due to the limited spectral bands in the vocoding process imposed by CI speech processing, listeners with CIs have partial access to fine-grained acoustic information in speech (Cleary, Pisoni, & Kirk, 2005; Houston, Pisoni, Kirk, Ying, & Miyamoto, 2003; Houston et al., 2012; Svirsky, 2017). Lack of access to these fine-grained acoustic cues such as pitch and low-frequency voice harmonics (Chatterjee & Peng, 2008; Deroche, Kulkarni, Christensen, Limb, & Chatterjee, 2016; Fu, Chinchilla, & Galvin, 2004; Gaudrain & Baskent, 2018; McDermott, 2004), as well as temporal fine structure cues from individual harmonics (Carlyon, Deeks, & McKay, 2010) may disrupt the ability of listeners with CIs to receive information contributing to signaling variations in talkers’ voice quality.
Listeners utilize acoustic information in snippets of talker’s voice during vowel phonation to draw inferences about talkers’ voice quality (e.g., Gerratt, J. Kreiman, 2016) and identity (Adams, 2009; Ladefoged & Broadbent, 1957; Purnell et al., 1999; Thomas, 2007). Acoustic information derived from sustained phonation of a vowel sound may reflect structural and functional properties of vocal folds vibrations (static and/or dynamic) and carries important information about identity and quality of talkers’ voice (Ladefoged & Broadbent, 1957; Morrison & Assmann, 2012; Winn, Chatterjee, & Idsardi, 2012). For example, fundamental frequency ($F_0$) and low-frequency harmonic components (e.g., $H1$-$H2$, $H1$-$A1$) are acoustic cues that reflect changes in the quality of talkers’ voice (e.g., breathy, strained, rough, etc.), and thus contribute to recognition of talkers’ identity (Chatterjee & Peng, 2008; Deroche et al., 2016; Fu et al., 2004; Gaudrain & Baskent, 2018; Gelfer & Bennett, 2013; McDermott, 2004; Van Dommelen, 1987). A similar role in talker recognition was shown for pitch dynamics (i.e., intonational patterns; Abberton & Fourcin, 1978). Chang et al., (2006) showed that the voice pitch cue in the low-frequency spectrum and the temporal envelop cues are important for segregation of the target voice from the competing voice and performing a robust recognition of speech. These acoustic cues, however, are poorly perceived by CI users due to the vocoding process in CI speech processing that involves a limited number of frequency channels (Carlyon et al., 2010; Laneau, Wouters, & Moonen, 2004). This was observed in prelingually deafened children with CIs who showed poorer performance compared to their NH peers in distinguishing between talkers with similar voices (Cleary & Pisoni, 2002). Low-numbered resolved harmonics are not faithfully transmitted via CI device, which negatively affects the performance of CI users in speaker identification and recognition (Gaudrain & Baskent, 2018), perception of prosodic information (Davidson, Geers, Uchanski, & Firszt, 2019; Deroche et al., 2016), and speech understanding in multi-talker
Talakers’ voice quality through cochlear implant (Rosen, Souza, Ekelund, & Majeed, 2013; Stickney, Zeng, Litovsky, & Assmann, 2004). Taken together, these prior findings indicate that some aspects of voice quality cues may be degraded through CI processing. Such findings may represent a combination of altered input at peripheral and central auditory pathways for sound and language processing. We currently have limited evidence about how CI speech processing may be detrimental to acoustic information relevant to talkers’ voice quality variations. Degradation of this information potentially disrupts the bridge between recognizing talkers’ voice and effective processing of their speech.

E. Present study

The present study analyzed voices from two groups of talkers with normal and disordered voice qualities to investigate the effect of simulated cochlear-implant speech processing and the number of spectral channels on the acoustic information involved in distinguishing between these two classes of voice qualities. Noise-vocoder was used to simulate the limited resolution of CIs in representation of acoustic information relevant to voice quality distinction. Voices spoken by individuals with voice disorders (e.g., vocal fold polyps, nodules, etc.) represent an instance where voice quality is abnormally altered compared to the voice of individuals with healthy vocal folds function (Arjmandi & Pooyan, 2012; Arjmandi, Pooyan, Mikaili, Vali, & Moqarehzadeh, 2011; Ghasemzadeh & Arjmandi, 2019; Umapathy, Krishnan, Parsa, & Jamieson, 2005). Abnormal physiological, neurological, and/or functional changes in vocal folds affect talkers’ voice quality, which in turn would be reflected as certain variations in the properties of talkers’ voice spectrum (Arjmandi & Pooyan, 2012; Behroozmand & Almasganj, 2007; Eskenazi et al., 1990; Fukazawa, el-Assuooty, & Honjo, 1988; Hammarberg, Fritzell, Gaufin, Sundberg & Wedin, 1980; Hanson, 1997; Klatt & Klatt, 1990; Kreiman, Gerratt, & Berke, 1994; Kreiman, Gerratt, & Precoda, 1990; Krom, 1995; O’Leidhin & Murphy, 2005; Sasaki & Okamura, 1984; Yumoto, Gould, & Baer,
These abnormalities, that signal the distinction between disordered and normal voice qualities, generally appear as time-varying noise-like patterns in different ranges of low-, mid- and high-frequency bands of voiced signals (e.g., Arjmandi & Pooyan, 2012; Behroozmand & Almasganj, 2007; Fukazawa, el-Assuooty, & Honjo, 1988; Hammarberg, Fritzell, Gaufin, Sundberg, & Wedin, 1980; Hanson, 1997; Klatt & Klatt, 1990; O’Leidhin & Murphy, 2005; Sasaki & Okamura, 1984; Yumoto, Gould, & Baer, 1982). Spectral changes in the vowel spectrum have shown to be important in monitoring variations in talkers’ voice quality such as breathiness, laryngealization, harshness, and nasality (Gerratt, Kreiman, 2016; Coleman, 1971; Klatt & Klatt, 1990; Park et al., 2016). In the present study, we investigated the variations in the magnitude spectrum of normal and disordered voices to understand how CI speech processing may impact the distinctive patterns of the voice quality. We further quantified the acoustic distance between normal and disordered voices as a function of the number of spectral channels to study the effect of the number of spectral channels on the degradation of voice quality information.

II. MATERIALS

A. Voice samples

The voice samples in this study were sustained vowel sounds /a/ from the voice disorders database model 4337, version 1.03 (Kay Elemetrics Corporation, Lincoln Park, NJ), developed by Massachusetts Eye and Ear Infirmary (MEEI), Voice and Speech Lab. Two groups of participants who were diagnosed as having either normal or disordered voices, were asked to sustain the vowel /a/ while their voices were recorded at a sampling frequency of 44.1 kHz with 16-bit resolution. All the analyzed voice segments were 1-second long, extracted from the middle of each excerpt to deal with the length difference between normal and disordered voice samples, as well as the transient patterns during the onset and offset of phonations. The vowel sounds from 293
individuals were analyzed where 53 talkers had normal voice quality (21 Males) and the remaining 240 talkers (96 Males) were diagnosed with one or multiple voice disorders. The voice disorders were the result of abnormal physiological, neurological, and/or functional changes that affected the regular function of talkers’ vocal folds.

**B. Creation of noise-vocoded voice samples**

The CI-simulated versions of the unprocessed voice stimuli were created using a noise-excited envelope vocoder in AngelSim™ Cochlear Implant and Hearing Loss Simulator (Fu, 2019; Emily Shannon Fu Foundation, www.tigerspeech.com), which follows the CI-simulated vocoding process in prior studies (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). The process involved dividing each voice spectrum into a variable number of logarithmically-spaced frequency bands between absolute lower and higher-frequency of 200 Hz and 7000 Hz (24 dB/Octave analysis filter slopes), corresponding to the frequency-place map simulated by Greenwood function (Greenwood, 1990). These frequency limits approximate the corner frequencies in the Cochlear Nucleus speech processors (Crew & Galvin, 2012; Winn & Litovsky, 2015). The amplitude envelope of each signal, obtained from filtering the voice spectrum under each sub-band, was captured using half-wave rectification and a low-pass filter with a cut-off frequency of 160 Hz and filter slope of 24 dB/oct to simulate the performance of the average CI listeners in envelope discrimination (Chatterjee & Oberzut, 2011; Chatterjee & Peng, 2008). The extracted amplitude envelopes were then used to modulate band-pass filtered white-noise carrier signals, which were filtered using a similar analysis filter. The final noise-vocoded version of each voice stimulus was created by summing amplitude-modulated signals. This process replaces fine spectro-temporal structures in voice signal with noise while preserving most of the course-grained temporal structures. The quality of CI-simulated voice is a function of the number of spectral
channels in the vocoder (e.g., Friesen, Shannon, Baskent, & Wang, 2001; Xu, Thompson, & Pfingst, 2005). The noise-excited envelope vocoder was used in AngelSim software to process unprocessed/natural voices and create their noise-vocoded versions with 4-, 8-, 12-, 16-, 22-, and 32-channel. Therefore, the simulated cochlear implant voices were created with seven levels of spectral degradation (unprocessed, 4-, 8-, 12-, 16-, 22-, and 32-channels). The choice of the number of spectral channels was made to simulate a wide range of spectral degradation and their corresponding perceived difficulty in speech processed through CIs (Shannon, Fu, & Galvin, 2004), as well as to cover the current set-up of between 12-24 active channels in cochlear implant devices. Considering the assumption made about the relationship between electrical spread in the cochlea and acoustical filter slope (Bingabr, Espinoza-Varas, & Loizou, 2008; Oxenham & Kreft, 2014), the selected filter slope of 24 dB/Oct is in the highest range of steepness provided by current CI technology (filter slope varies between 8 and 24 dB/oct), corresponding to the minimum channel interaction available in the current CI devices.

III. METHODS

A. Analysis of voice spectra

We first explored the average spectra of two groups of normal and disordered voice quality to understand how noise-vocoder and the number of spectral channels affect the distinctive features of two classes of voice qualities within different frequency regions. The average spectrum of voice signals derived over the spectrum of all voice samples from each class of voice quality (i.e., normal or disordered) were estimated using linear predictive coding (LPC) with 12th order (Rabiner & Schafer, 1978). These average spectra were comparatively used to investigate how the distribution of energy of the voice signal is changed over various frequency regions due to CI noise vocoding process. In fact, the variations in the characteristics of the average voice spectra were investigated
under seven levels of spectral degradations (unprocessed, 32, 22, 16, 12, 8, and 4-channels noise vocoder) to identify how CI filtering process affects the acoustic information signaling distinction between two classes of normal and disordered voice quality.

**B. Using MFCC features to characterize acoustic information relevant to voice quality variation**

We used mel-frequency cepstral coefficients (MFCCs) to characterize variations in voice quality for normal and disordered voices and their corresponding CI-simulated versions. The CI-simulated voices were created at 6 levels of spectral degradations (i.e., number of spectral channels). These features were inspired by human auditory processing and were originally developed to approximate the filtering structure and frequency resolution of the human auditory system (Fant, 1973; Hunt, Lennig, & Mermeletein, 1980; Davis and Mermelstein, 1980; Shaneh & Taheri, 2009; Stevens, Volkmann, & Newman, 1937). These acoustic features have been effective in representation of the variations in voice quality, particularly in discrimination between normal and disordered voices (Ali, Alsulaiman, Muhammad, Elamvazuthi, & Mesallam, 2013; Dibazar, Narayanad, & Berger, 2002; Firdos & Umarani, 2016; Godino-Llorente, Gomez-Vilda, & Blanco-Velasco, 2006; Panek, Skalski, Gajda, & Tadeusiewicz, 2015). Furthermore, features that are developed based on speech production behaviors such as $F_0$ and the first and second harmonics ($H1$ and $H2$) are either heavily degraded or even absent in the CI-simulated voices. Using MFCC features, we were able to deal with such methodological challenges.

Fig. 1A shows the schematic of the approach for calculating MFCCs for samples of normal and disordered voices. To calculate MFCCs, each vowel stimulus /a/ (i.e., $S_n$ or $S_d$) was first segmented into frames of 30 ms with a frame shift of 15 ms. Here the $i$ and $j$ indicate the index of voice stimuli for normal and disordered voices, respectively ($i = \{1,2,3, \ldots,53\}$ and $j = \{1,2,3,$
A Hamming window was then applied to each frame to decrease the effect of sidelobes for better frequency-selective analysis (Rabiner & Schafer, 1978). The power spectrum of each frame was calculated based on Fast Fourier transform (FFT) analysis. Then, the mel-filterbank were generated and applied to the voice power spectrum. MFCCs were eventually derived by calculating the DCT of the logarithm of all filterbank energies (Rabiner & Schafer, 1978). Eventually, the first twelve components were preserved as MFCC features for each frame of a voice stimulus (i.e., $MFCC_{ni}$ or $MFCC_{dj}$). Each voice stimulus was eventually represented by an MFCC matrix with the size of $F \times 12$, where $F$ indicates the number of frames in each voice stimulus. The same procedure was performed to calculate MFCC matrices for CI-simulated versions of natural/unprocessed normal and disordered voices, as shown by dashed lines in Fig. 1A.

C. Acoustic distance quantification using Mahalanobis distance measure

To examine the acoustic distance between two classes of normal and disordered voices as a function of the level of spectral degradation (imposed by the variable number of spectral channels in the CI-simulated voices), we calculated Mahalanobis distance (MD) on MFCC features. MD is a distance measure, which calculates the distance between two or more classes at a multidimensional feature space (Arjmandi et al., 2018; Maesschalck & Massart, 2000; Masnan et al., 2015; Xiang, Nie, & Zhang, 2008). This multivariate statistical approach uses two feature matrices (or vectors) from two separate classes to evaluate the extent that the two classes are distinguished, after sphering the distance matrix between two classes using the average covariance matrix of the per-class centered data (Maesschalck & Massart, 2000; Masnan et al., 2015). Hence, a relatively greater MD value for a condition (e.g., unprocessed) means a relatively larger distance...
between two classes of normal and disordered voice qualities with a relatively lower between-class overlap for that condition compared to other conditions (e.g., 32-channel noise-vocoded voice). A relatively larger distance indicates the presence of more discriminative acoustic information relevant to the distinction between two classes of voice qualities for a condition compared to other conditions.

As shown in Fig. 1B, the acoustic properties of each voice stimulus (i.e., $S_{ni}$ or $S_{dj}$) were characterized by a single, time-averaged MFCC vector (i.e., $\mu_{MFCC_{ni}}$ or $\mu_{MFCC_{dj}}$), derived by averaging MFCCs across frames of each voice stimulus. These time-averaged MFCC features have been shown to successfully represent the unique spectral characteristics of a sound (Mckinney & Breebaart, 2003; Davis and Mermelstein, 1980; Terasawa, Slaney, & Berger, 2005). Therefore, 53 normal voices were presented by 53 12-dimensional average MFCC vectors, constructing a $53 \times 12$ feature matrix. Likewise, 240 disordered voices were characterized by 240 12-dimensional average MFCC vectors, leading to a $240 \times 12$ feature matrix. The acoustic distance between normal and disordered voices was eventually measured by calculating MD between these two matrices of MFCCs (Fig. 1B). The same procedure was followed for simulated versions of the unprocessed normal and the unprocessed disordered voice stimuli (dashed lines in Fig. 1B), leading to seven values of MD corresponding to seven levels of spectral degradation from the unprocessed to 4-channel noise-vocoded voices. The calculated MDs of seven levels of spectral degradation were comparatively examined to identify the extent to which the spectral information involved in distinguishing voices with normal and disordered qualities were affected by CI noise vocoding, as well as the number of channels in the CI noise-vocoding process.
D. Acoustic distance quantification based on the classification accuracy of support vector machines (SVMs)

SVMs have been frequently used as a successful classification method for various classification purposes, including classification of normal and disordered voice qualities (Akbari & Arjmandi, 2015; Arjmandi et al., 2011; Ghasemzadeh, Khass, Arjmandi, & Pooyan, 2015; Ghasemzadeh & Arjmandi, 2019; Arjmandi, Pooyan, Mohammadnejad, & Vali, 2010; Umapathy, Rachel, & Thulasi, 2018). An application of SVM classifier is the evaluation of features in distinguishing between two classes, where the classification accuracy of SVMs classifier is used as a criterion for feature evaluation (Heijden, Ferdinand, Ridder, & Tax, 2005). We took advantage of this property of SVM to examine the effect of CI speech processing and the number of channels on the acoustic distinctiveness between two classes of normal and disordered voices as a complementary analysis to MD. Higher classification accuracy between two classes indicates that there was more distinctive acoustic information with respect to class separation.

As illustrated in Figure 1C, a 5-fold cross-validation analysis was performed to train and then test an SVM classifier on its classification accuracy in distinguishing between normal and disordered voice qualities at seven levels of spectral degradation (unprocessed, 32-, 22-, 16-, 12-, 8-, and 4-channel). Two feature matrices of 53x12 and 240x12 MFCC features from normal and disordered classes were entered into the SVM classifier to be used in training and testing phases as executed through the 5-fold cross-validation procedure (Kohavi, 1995; Reilly, Moran, & Lacy, 2004). The output of the SVM classifier was the mean SVM classification accuracy over classification accuracies, obtained from five repetitions of cross-validation. The average classification accuracies at six levels of spectral degradation were examined with reference to that of the unprocessed condition (as baseline performance) to understand the extent of degradation
imposed by CI noise vocoding on acoustic information involved in voice quality distinction. The radial basis function (RBF) kernel was used in SVM classifier. The parameters of RBF kernel and the regularization parameter ($\xi$) of SVM were set to their default values in Matlab. Compared to MD, the training phase in SVM allowed us to simulate the effect of exposure to voice samples from two classes of voice qualities in their acoustic distinction in a computational fashion. The training phase in SVM is expected to benefit from seen data in a supervised fashion to associate degraded patterns in CI-simulated voices to their corresponding class of voice quality (i.e., normal or disordered). This effect is basically very important in listeners with CIs as studies have shown that the central auditory system plays a significant role in decoding the degraded acoustic patterns at the outputs of CI electrodes in CI users (Basura, Hu, Juan, Tessier, & Kovelman, 2018; Fallon, Irvine, & Shepherd, 2008; Houston et al., 2012).

IV. RESULTS

A. Effects of CI noise vocoding on spectral information relevant to voice quality distinction

We first examined the vowel /a/ spectra from normal and disordered voices under seven spectral degradation conditions. Fig. 2 shows the average magnitude spectra of the two groups of voices with normal (blue or dark gray) and disordered (orange or light gray) qualities across all voice samples. The standard deviations of magnitude spectra are also shown as a shaded area across the average lines. The average voice spectra are selectively shown for unprocessed/natural (panel A) and simulated cochlear implant voices with 16- (panel B) and 4-channel (panel C) spectral resolution. These spectra are computed by averaging individual frequency spectrum over all voice samples from a class of voice quality. Overall, these plots demonstrate the detrimental effect of CI noise vocoding process on spectral information that could signal differences in talkers’ voice quality. The patterns of variation in average spectral energy of disordered voices compared to
normal voices at different frequency sub-bands reflect voice quality variations, caused by a wide range of physiological, neurological, and/or functional voice disorders.

The difference between the average magnitude spectrum for normal voices (the blue or dark gray line) and that of disordered voices (the orange or light gray line) in Fig. 1A reveals distinctive patterns of spectral energy within low-, mid-, and high-frequency regions. For unprocessed/natural condition, a peak in the frequency regions between 1 and 2 kHz distinguishes average spectrum of voices with normal quality from the average spectrum of disordered voices. These degraded low-frequency patterns in the disordered voice spectrum may be signs of partial closure of vocal folds. These differences in the spectral level in frequency bands covering the first formant (e.g., breakdown in formant structure) can be associated with breathy voice quality reported in some voice disorders (Kitzing & Åkerlund, 1993; Krom, 1995; Rontal, Rontal, & Rolnick, 1975; Thomas, 2008; Wolfe & Bacon, 1971). The relative reduction in low-numbered harmonic components is also visible in low-frequency regions which is due probably to irregular vibratory patterns of vocal folds and hoarse voices in disordered voice quality (Fex, Fex, Shiromoto, & Hirano, 1994; Roy & Leeper, 1993; Thomas, 2008; Yanagihara, 1967). A relatively higher level of energy in mid-frequency bands (~4.7 KHz-12.4 kHz) is evident in the average spectrum of disordered voice qualities compared to that of the normal group, which potentially signals the presence of high degree of breathiness in disordered voice samples and an increase in the level of the turbulence noise components in the vocal excitation signal (Askenfelt & Hammarberg, 1986; Fukazawa et al., 1988; Hanson, 1997; O’Leidhin & Murphy, 2005). The presence of a wide-band noise in this frequency region (i.e., between ~ 5 kHz and ~12 kHz) in the average spectrum of disordered voices may be attributed to a rough voice quality in disordered voices (Krom, 1995). Finally, the higher degree of variability in voice spectrum across the average
spectrum in disordered class (the wider orange (light gray) shaded area) compared to the one for
normal voice quality (blue (dark gray) shaded area) can be attributed to variable levels of
abnormality in vocal folds vibrations in disordered voices, as well as the variety of the analyzed
voice disorders.

More importantly, comparing the average spectrum of two groups of voices with normal
and disordered qualities among three levels of spectral degradation in Fig. 2 (unprocessed in panel
(A), 16-channel in panel (B), and 22-channel in panel (C)) demonstrates that CI noise vocoding
substantially degrades acoustic information involved in voice quality distinction. In general, the
noise vocoding process in CI speech processing caused major loss of acoustic information
distinctive of talkers’ voice quality at low-, mid-, and high-frequency ranges of their voice spectra.
The detrimental effect of CI noise-vocoder increases as the number of channels decreases to the
extent that spectra of two classes of voice qualities become almost visually indistinguishable at 4-
channel CI-simulated voices, highlighting the substantially detrimental effect of low spectral
resolution in CI speech processing on discarding voice quality-related acoustic information. As the
spectra of normal and disordered voices for 16- and 4-channel noise-vocoded voices suggest, a
large portion of spectral information at low frequency regions is discarded due to the filtering
process in CI. This spectral region is particularly important for the perception of voice quality
variations as it is where the low-numbered harmonics are located. CI listeners do not have access
to low-numbered resolved harmonics, which are important for robust pitch perception (Bernstein
& Oxenham, 2003; Smurzynski, 1990). The noise level elevated in the mid-frequency band as the
number of spectral channels decreases in the CI-simulated voices, likely leading to lack of access
to acoustic cues relevant to variations in talkers’ voice quality in listeners with CIs.

Fig. 2 about here
B. Effects of CI noise vocoding on acoustic information distinctive of talkers’ voice quality

As discussed in the Method section, MFCCs for each voice stimulus were calculated to characterize talkers’ voice quality and measure the effect of CI noise-vocoding on acoustic information relevant to voice quality distinction. Fig. 3 illustrates an example of the filterbank of mel-spaced triangular filters through which voice spectrum of each normal or disordered voice was passed to characterize the energy variations at different frequency sub-bands.

Fig. 4 shows the calculated *Mahalanobis distance* between MFCCs of normal and disordered voices as a function of different levels of spectral degradation, corresponding to change in the number of spectral channels in CI-simulated voices. This figure illustrates that the acoustic distance between voices with normal and disordered qualities decreased due to the CI noise vocoding process, suggesting that CI speech processing is detrimental to voice quality-related acoustic information. On average, there was an approximately 33% decline in MD due to CI noise-vocoding process when comparing the MD at unprocessed condition (the top dashed line in Fig. 4) with the average MD derived across six levels of spectral degradation (the middle dotted line in Fig. 4). This large decline in acoustic distance between normal and disordered voice qualities suggests that the CI noise-vocoder potentially discards an important portion of acoustic information responsible for signaling variation in talkers’ voice qualities. An unexpected pattern was the increase in MD as the number of spectral channels decreased from 32 channels to 22, 16, and 12 channels. Our visual investigation of normal and disordered voice spectra showed that noise vocoding interestingly resulted in more distinctive patterns of spectral energy between normal and disordered voice qualities as the number of spectral channels changed from 32 to 12 channels. This pattern was particularly noticeable in low frequency regions where mel-filterbank is more sensitive
to variations in voice spectrum because of its narrower filters, as compared with high-frequency regions with wider filters (see Fig. 3). This unexpected pattern suggests that increasing spectral channels in CI noise-vocoder does not necessarily mitigate the information loss relevant to voice quality distinction. The range of decrease in acoustic distance (i.e., MD) due to noise vocoding relative to acoustic distance in the unprocessed condition was between ~64% for 4-channel CI-simulated voices and ~19% for 12-channel CI-simulated voice, which was still relatively high.

Fig. 5 shows the results of normal-vs-disordered SVM classification accuracy for seven levels of spectral resolution from the unprocessed condition to the highly spectrally-degraded CI-simulated voices created by 4-channel noise-vocoder. Results from SVM classification supports the general trend displayed by MD on the effect of cochlear implant speech processing on acoustic information involved in normal-vs-disordered voice distinction. However, there was an approximately 8% decline in the accuracy of SVM in classification of normal and disordered voice qualities between the unprocessed condition and the average accuracy obtained across six levels of spectral degradation. This decline is much smaller than the ~33% decline measured by MD, which potentially highlights the key role of classifier exposure to seen data during training phase. Classification accuracy in Fig. 5 shows three categories of performance between 80-85%, 85-90%, and 90-95%. These simulated results suggest that the current CI technology falls within the second category in terms of the number of spectral channels (12 channels in MED-EL devices, 16 channels in Advanced Bionics devices, and 22 channels in Cochlear), where the classification accuracy is still at least 5% below the one for the unprocessed condition. It is notable that classification accuracies for CI-simulated conditions fall within a good performance range (80% to 90%), even for a highly degraded CI-simulated voice with 4 spectral channels. We speculate that this
difference between SVM and *Mahalanobis distance* in measuring voice quality-related acoustic distinction is because of the exposure phenomenon simulated by SVM as being trained on a subset of data in a supervised fashion. Another explanation could be related to the calculation of MD, which assumes that features have multivariate normal distribution, which might not be necessarily valid for the MFCC features in this study.

Fig. 5 about here <<<<<<<<

V. DISCUSSION

This study investigated how CI speech processing affects acoustic information involved in signaling variations in talkers’ voice quality. We analyzed vowel sounds /a/ spoken by 53 talkers with normal voice and 240 talkers with disordered voice qualities at seven levels of spectral degradation, simulated by CI noise-excited envelope vocoder, to examine the effect of CI speech processing on acoustic information that distinguishes talkers’ voice quality. To our knowledge, this is the first study that examines the effect of CI speech processing on acoustic information involved in conveying variations in talkers’ voice qualities. Overall, the results of the current study highlighted the detrimental effect of CI noise vocoding process on acoustic information that signals voice quality contrast. This unfaithful transmission of acoustic information relevant to voice quality distinction through CIs may negatively impact the performance of CI listeners in identification and recognition of talkers’ voice and, thus comprehension of their speech.

Our investigation of vowel /a/ spectra within different frequency sub-bands across two groups of normal and disordered voice qualities showed that simulated CI processing, based on a noise-vocoder, has a detrimental impact on acoustic information signaling changes in talkers’ voice quality. The CI noise-vocoder processing substantially degraded spectral information in low- (<2 kHz), mid- (~4-12 kHz), and high-frequency ranges (>12 kHz) that could contribute to voice
quality distinction. The discriminative spectral information under these frequency regions signals various degrees of distinctive acoustic information that listeners may utilize to perceive fine variations in talkers’ voice quality (Eskenazi et al., 1990; Hillenbrand, Cleveland, & Erickson, 1994; Moisik, 2013; Park et al., 2016; Podesva, 2007; Sicoli, 2010). Our results suggest that CI speech processing substantially degrades this distinctive spectral information, likely leading to a degraded perception of talkers’ voice quality variations. These patterns of loss in voice quality-related information due to CI voice processing suggest that listeners with CIs potentially do not receive a large portion of acoustic information signaling changes in talkers’ voice quality due to the partial transformation of fine-grained spectral structures.

We further measured the spectral distance between voices with normal and disordered qualities by first characterizing their vowels’ spectral variations using MFCC features and then calculating the distance between MFCC features using MD. The MDs between normal and disordered vowel sounds were examined at different levels of spectral degradation (i.e., unprocessed, 32, 22, 16, 12, 8, and 4-channel CI noise-vocoder processing) to identify how simulated CI speech processing affects the acoustic distance between voices with normal and disordered qualities. We further examined this effect as a function of the number of spectral channels in the noise-vocoder. The results showed a large decrease in acoustic distance between normal and disordered voice qualities because of CI speech processing, highlighting the loss of acoustic information related to talkers’ voice quality through CI. Therefore, listeners with CIs potentially face difficulties compared to listeners with NH in incorporating talkers’ voice quality information to construct the corresponding mental representation for identification and recognition of talkers’ voice and processing their speech.
The results of SVM accuracy in classification between normal and disordered voice qualities corroborated the detrimental effect of CI noise vocoding on acoustic information involved in voice quality distinction. However, the CI noise-vocoder processing and the number of channels in the CI noise-vocoder resulted in lower degrees of drop based on SVM classification compared to the amount of decline quantified by MD. These results also simulated the effect of exposure/learning in making sense of the degraded voice signals for voice quality distinction. The results of SVM classification, as executed through 5-fold cross-validation procedure, suggested a good performance of higher than 80%, even in the highly degraded condition of 4 spectral channels. The average drop of ~ 7% in normal-vs-disordered voice classification between unprocessed and CI-simulated conditions highlights the detrimental effect of CI speech processing to acoustic information relevant to recognition of talkers’ voice.

These results can be interpreted in the context of perception of voice cues involved in talker recognition, as well as speech processing. The observed lack of faithful transmission of acoustic information, that is more or less related to various perceptual attributes of voice quality (e.g., breathiness, harshness, creakiness, and nasality), suggests that CI listeners may perform poorer than their peers with NH in processing segmental and suprasegmental information for speech comprehension (Dicacio, 2009; Dilley et al., 1996; Dilley et al., 2016; Garellek & Keating, 2011; Gordon, 2001; Gordon & Ladefoged, 2001; Henton, 1986; Ogden, 2001; Redi & Shattuck-Hufnagel, 2001), as well as in recognition of talkers’ gender (Gussenhoven, 2004; Ohala, 1983; Puts, Hodges, Cárdenas, & Gaulin, 2007), race (Alim, 2004; Irwin, 1977; Moisik, 2013; Thomas & Reaser, 2004) and social and cultural class (Esling, 1978; Rilliard et al., 2009; Sicoli, 2007; Stross, 2013; Stuart-Smith, 1999). Our investigation of normal and disordered voice qualities suggests that CI noise vocoding substantially degrades spectral properties signaling voice quality...
variations (Dicanio, 2009; Garellek & Keating, 2011), which probably negatively impact CI listeners’ access and learning talker-specific information as an important skill for robust speech recognition (Johnson, 2005; Kleinschmidt & Jaeger, 2015; Pisoni, 1992). Results from prior studies demonstrated that listeners with CIs do not have access to low-numbered harmonics for robust perception of $F_0$, leading to poor performance in talker identification and discrimination (Gaudrain & Baskent, 2018), prosody perception elicited by dynamic pitch (Deroche et al., 2016), and speech recognition in complex listening conditions such as multi-talker situations (Rosen et al., 2013; Stickney, Assmann, Chang, & Zeng, 2007; Stickney et al., 2004). In addition, listeners with NH may incorporate other cues such as vocal-tract length (VTL) and formant frequencies in constructing talkers’ voice quality to distinguish between talkers, cues that are poorly perceived by listeners with CIs (Gaudrain & Baskent, 2018). Our results provide further evidence in explaining the poor performance of listeners with CIs in perception and effective use of talkers’ voice cues (Başkent, Luckmann, Ceha, Gaudrain, & Tamati, 2018; Gaudrain & Baskent, 2018; Mehta, Lu, & Oxenham, 2020; Mehta & Oxenham, 2017; Moore & Carlyon, 2005; Stickney et al., 2007) by showing that an important portion of this acoustic information is discarded by cochlear implant speech processing.

There are multiple limitations in the current CI devices including the number of channels in the vocoder, which restricts spectral and temporal resolution of CI devices in representation of speech. Our results highlight the need for improving CI speech processing strategies to assure that acoustic cues related to voice quality are faithfullly transferred through CIs. Developing more effective strategies requires researchers to evaluate the mechanisms underlying encoding spectral and temporal cues responsible for representing voice quality measures. Therefore, further studies are required to understand how listeners with CIs perceive acoustic cues related to voice quality.
variations and how possible loss of information at this level may impact their ability to identify talkers and process their speech. Tamati et al., (2017) found that listeners with CIs perform poorer than their NH peers in speech recognition when there is large talker variability. They also showed that CI users experience difficulties in the recognition of talkers’ voices and accents while their performance were also largely variable compared to listeners with NH. Results from the present study provide further evidence that listeners with CIs may not have access to voice quality cues for robust identification of talkers. This lack of access to the voice quality cues may negatively impact CI listeners’ ability to overcome talker variability for successful speech perception.

Despite the limitations of CI in robust and reliable transformation of speech, Vongpaisal et al. (2010) showed that children with CIs are able to develop models of talker identity, which may reflect the important role of neural plasticity and more powerful speech processing at higher cortical levels for auditory processing and language development. As simulated by SVM, there might be a large effect of exposure or training that can improve the performance of CI listeners in distinction between various voice qualities. In fact, this phenomenon can be logically expanded to how CI listeners may use the information in the voice delivered through CI at higher levels of speech processing and language learning to compensate for the lack of various acoustic cues such as those related to the perception of talkers’ voice quality (Moore & Shannon, 2009). Speech recognition of children with CIs significantly improved as they had more experience in listening to speech through a CI device (Brown et al., 2004; Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Miyamoto, Osberger, & Kessler, 1996; Tyler et al., 2000). Another factor that is not modeled in our study is the effect of linguistic and contextual cues in continuous speech that listeners with CIs can incorporate to infer talkers’ voice quality for talker recognition and language processing. The significant effect of these cues on sentence recognition was shown in listeners
with CIs (Geers, 2002; Meyer & Svirsky, 2000). Despite these contextual effects, having access
to acoustic information relevant to talkers’ voice quality is still critical for speech processing and
language development (Başkent et al., 2018; Gaudrain & Baskent, 2018), particularly in complex
listening conditions such as speech recognition in multi-talker scenarios (Rosen et al., 2013;
Stickney et al., 2007, 2004).

The present study had some limitations that should be considered while interpreting the
findings. Although studies based on CI-simulated speech are advantageous in general, these
findings should be viewed as the general trend rather than the actual performance of CI listeners
in perception of talkers’ voice quality. Furthermore, characterization of voice quality based on
MFCCs features might not completely reflect the normal hearing system in perception of voice
quality variations as shown by recent studies (Anand, Kopf, Shrivastav, & Eddins, 2019; Eddins,
Anand, Lang, & Shrivastav, 2020). It is also worth mentioning that listeners may incorporate
segmental and suprasegmental cues at word and/or sentence levels for recognition of talkers’ voice
quality rather than merely relying on spectral variations of vowel sounds. Regardless of these
limitations, the present study provided new evidence showing that acoustic information involved
in distinguishing talkers’ voice quality is substantially degraded in CI-simulated voices. Our results
suggest that listeners who use CIs may have great difficulties incorporating voice quality cues for
talkers’ voice recognition. The poor spectral resolution provided by cochlear implant device to CI
listeners negatively impacts acoustic cues involved in voice quality transmission, leading to
subsequent poor perception of talkers’ voice quality in listeners with CIs. This degraded
transmission of acoustic cues relevant to talkers’ voice quality is particularly detrimental for
speech processing and auditory scene analysis in prelingually deaf children who heavily rely on
the bottom-up sensory information in speech to develop mental representations specific to talkers’
voice. Future perceptual studies will determine which specific acoustic cues relevant to talkers’ voice quality are not faithfully transmitted through cochlear implants. The findings from the current study underscore the need in two directions: (a) the need for examining the current signal processing strategies in CIs for their fidelity in passing voice quality cues and developing more advanced speech processing strategies in CI device to assure faithful transmission of these cues, and (b) the need for active use of multimodal (i.e., gesture, tactile, and visual) communicative behaviors to provide supportive cues for listeners with CIs in recognition of talkers’ voice, especially in pediatric CI users.

REFERENCES


Running title: Talkers’ voice quality through cochlear implant


Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic Correlates of Breathy Vocal


Talkers’ voice quality through cochlear implant


Shannon, Robert V., Fu, Q. J., & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Oto-
Laryngologica, Supplement, 124(552), 50–54. https://doi.org/10.1080/03655230410017562


1014 1081–1091. https://doi.org/10.1121/1.1772399

Society, 42*(2), 139–162. https://doi.org/10.1017/S004740451300002X

British Isles, 203–222.*


recognition by cochlear implant users. *The Journal of the Acoustical Society of America,

of ICAD 05-Eleventh Meeting Of the International Conference on Auditory Display, 6–9.*


Thomas, E. R., & Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of
African American and European American voices. *Journal of Sociolinguistics, 8*(1), 54–87.
https://doi.org/10.1111/j.1467-9841.2004.00251.x

Aggressiveness of the growl-like timbre: Acoustic characteristics, musical implications, and


Fig. 1. (Color online) Schematic diagram of the approach used in the current study to (A) characterize acoustic properties of normal and disordered voice stimuli based on MFCCs features, (B) evaluate the acoustic distance between voices with normal and disordered qualities at seven levels of spectral degradation from no degradation (unprocessed vowel sound) to CI-simulated voices with 4 channels based on calculating Mahalanobis distance between MFCCs matrices of the two classes, (C) evaluate the acoustic distance between two classes of normal and disordered voice qualities as a function of levels of spectral degradation based on the classification accuracy derived from applying 5-fold validation to SVM classifier. The SVM classifiers were trained and tested on MFCCs features obtained from voice stimuli from two classes of normal and disordered voice qualities. The dashed lines refer to the process for creating and analyzing the CI-simulated versions of the unprocessed voice stimuli. N in the “N-channel Simulator” block stands for the number of spectral channels in the CI-simulated noise vocoder. Components, blocks, and lines with blue (dark gray) color show the paths for processing voices with normal quality, whereas the components with orange (light gray) color show the paths for processing voices with disordered quality. $S_{ni}$ or $S_{dj}$ in panel (A) represent normal and disordered voice stimuli for subject number $i$ ($i=\{1,2,\ldots, 53\}$) and subject number $j$ ($j=\{1,2,\ldots, 240\}$), respectively. $\text{MFCCs}_{ni}$ and $\text{MFCCs}_{dj}$ are MFCC matrices extracted from frames of $S_{ni}$ and $S_{dj}$ signals. $\mu_{\text{MFCCs}_{ni}}$ and $\mu_{\text{MFCCs}_{dj}}$ in panel (B) and (C) are average 12-dimensional MFCCs obtained over frames of $\text{MFCCs}_{ni}$ and $\text{MFCCs}_{dj}$ matrices. $MD$ in panel (B) is the Mahalanobis distance calculated on average MFCCs matrices derived from voice stimuli, belonging to two classes of normal and disordered voice qualities. Panel (C) shows the procedure of 5-fold cross-validation in 5 iterations applied to SVM to evaluate the effect of spectral degradation in CI-simulated voices on the acoustic separation between two classes of normal and disordered voices. F1 to F5 stand for Fold 1 to Fold 5 where the entire feature...
data (i.e., MFCCs features) was split into 5 folds; 4 folds for training the SVM classifier and the remaining fold for testing the SVM. Dashed rectangles indicate the folds for testing the classifier and solid rectangles refer to those for training.

**Fig. 2.** (Color online) Average magnitude spectra of voice stimuli with normal (blue or dark gray lines) and disordered (orange or light gray lines) qualities for (A) unprocessed voices, (B) simulated cochlear implant voices with 16-channel, and (C) 4-channel in the CI noise-vocoder. The standard deviations of the magnitude spectra are also shown as the blue (dark gray) shaded area across the average blue line (dark gray) for the group of normal voices and as the orange (light gray) shaded area across the average orange line (light gray) for the group of disordered voices.

**Fig. 3.** (Color online) The filterbank of mel-spaced triangular filters (green or dark gray dotted lines) superimposed on average magnitude spectra of voices with normal (blue or dark gray line) and disordered (orange or light gray line) qualities. In this example, mel-filterbak contains 12 filters, which starts at 0 Hz and expands to 22.05 kHz, corresponding to half of the sampling frequency (44.1 kHz). The actual filterbank in the calculation of MFCC features was comprised of 32 mel filters.

**Fig. 4.** (Color online) Mahalanobis distance between two groups of voices with normal and disordered qualities as a function of spectral degradation in CI-simulated voices (i.e., number of spectral channels). The top, horizontal dashed line shows the MD derived from unprocessed voices and the dotted line in the middle shows the average of MDs across six levels of spectral degradation. The unprocessed condition is labeled as “NT”, which stands for natural/unprocessed stimuli.
Fig. 5. (Color online) The accuracy of SVM in classification between two groups of normal and disordered voices at seven levels of spectral degradation, corresponding to change in the number of noise-vocoder frequency channels (i.e., unprocessed, 32-, 22-, 16-, 12-, 8-, and 4-channels noise-vocoder). The unprocessed condition is labeled as “NT”, which stands for natural/unprocessed stimuli.