Effect of spectral normalization on different talker speech recognition by cochlear implant users

Chuping Liu[a)]
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089

John Galvin III
Department of Auditory Implants and Perception, House Ear Institute, 2100 West Third Street, Los Angeles, California 90057

Qian-Jie Fu
Department of Biomedical Engineering, University of Southern California, Los Angeles, California 90089 and Department of Auditory Implants and Perception, House Ear Institute, 2100 West Third Street, Los Angeles, California 90057

Shrikanth S. Narayanan
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089

(Received 25 January 2007; revised 12 February 2008; accepted 23 February 2008)

In cochlear implants (CIs), different talkers often produce different levels of speech understanding because of the spectrally distorted speech patterns provided by the implant device. A spectral normalization approach was used to transform the spectral characteristics of one talker to those of another talker. In Experiment 1, speech recognition with two talkers was measured in CI users, with and without spectral normalization. Results showed that the spectral normalization algorithm had small but significant effect on performance. In Experiment 2, the effects of spectral normalization were measured in CI users and normal-hearing (NH) subjects; a pitch-stretching technique was used to simulate six talkers with different fundamental frequencies and vocal tract configurations. NH baseline performance was nearly perfect with these pitch-shift transformations. For CI subjects, while there was considerable intersubject variability in performance with the different pitch-shift transformations, spectral normalization significantly improved the intelligibility of these simulated talkers. The results from Experiments 1 and 2 demonstrate that spectral normalization toward more-intelligible talkers significantly improved CI users' speech understanding with less-intelligible talkers. The results suggest that spectral normalization using optimal reference patterns for individual CI patients may compensate for some of the acoustic variability across talkers.

© *2008 Acoustical Society of America.* [DOI: 10.1121/1.2897047]

PACS number(s): 43.71.Gv, 43.64.Me, 43.72.Fx, 43.66.Ts [ARB] Pages: 2836–2847

I. INTRODUCTION

Normal hearing (NH) listeners are able to understand speech from a variety of talkers, despite differences in acoustic characteristics (e.g., voice pitch, speaking rate, accent, etc.). NH listeners are thought to use some form of "speaker normalization" to process speech from multiple talkers, thereby preserving the perceptual constancy of the linguistic message (Pisoni, 1993). Speaker normalization may affect processes at an early segmental acoustic-phonetic level (Verbrugge *et al.*, 1976; Assmann *et al.*, 1982), and is associated with some central processing cost, as reflected in the reduced speech performance as the number of talkers is increased (Mullennix *et al.*, 1989; Sommers *et al.*, 1994).

Despite the operation of such speaker normalization processes, speech intelligibility varies considerably across different talkers. Different talkers have been shown to produce different levels of speech intelligibility in NH listeners (e.g., Hood and Poole, 1980; Cox *et al.*, 1987). For example, Cox *et al.* (1987) studied the intelligibility of speech materials produced from three male and three female talkers in different listening conditions in NH listeners. Results indicated significant differences in intelligibility across talkers, even in listening environments that allowed for full intelligibility of everyday conversations. These cross-talker effects have also been observed in cochlear implant (CI) users. For example, Green *et al.* (2007) recently studied the effects of cross-talker differences on speech intelligibility in CI users and NH listeners listening to acoustic CI simulations. In their study, two groups of talkers (high or low intelligibility talkers) were established according to mean word error rates, based on previous data collected with NH listeners; each group consisted of one male adult, one female adult, and one female child talker. Results showed differences in intelligibility between the two talker groups, for a variety of listening conditions; talker group differences were maintained even when overall speech performance was reduced in the more difficult listening conditions. In CIs, speech patterns are represented by a limited number of spectral and temporal cues. In addition, electrically evoked speech patterns may be further dis-

[a)]Electronic mail: chupingl@usc.edu

torted due to the spectral mismatch between the input acoustic frequency and electrode place of stimulation. Previous CI acoustic simulation studies with NH listeners have shown differences in speech understanding for different talkers (e.g., Dorman *et al.*, 1997a; Fu and Shannon, 1999). While talker variability may not have been the main research focus, these studies suggest that the degree of spectral distortion may significantly affect intelligibility with different talkers. Thus, compared with NH listeners whose spectral resolution may better support perceptual normalization across talkers, CI users' speech recognition may be more susceptible to acoustic differences across talkers.

Although widely studied, the relation between speech intelligibility and the acoustic-phonetic properties for different talkers remains unclear. For example, while Bradlow *et al.* (1996) found no correlation between speaking rate and intelligibility in NH listeners, Bond and Moore (1994) found that, compared with more-intelligible talkers, less-intelligible talkers produced words and vowels with shorter durations. Besides speaking rate, other acoustic-phonetic correlates of intelligibility across talkers have been studied, e.g., fundamental frequency (F0; Bradlow *et al.*, 1996), amplitude of stressed vowels (Bond and Moore, 1994), and long-term average spectrum and consonant-vowel intensity ratio (Hazan and Markham, 2004). Typically, no single acoustic feature was able to explain the intelligibility difference across talkers. Hazan and Markham (2004) suggested that highly intelligible speech may depend on combinations of different acoustic-phonetic characteristics.

Some researchers have tried to improve speech intelligibility by compensating for differences along one acoustic dimension. For example, Luo and Fu (2005) studied an acoustic rescaling algorithm to normalize the formant space across talkers; the algorithm was evaluated in NH subjects listening to acoustic CI simulations. In their study, mean third formant frequency (F3) values (across vowels) were calculated for each talker in the stimulus set. The ratio between the mean F3 value for each talker and the reference talker (the talker that produced the best vowel recognition for each subject) was used to adjust the analysis filter bank of an acoustic CI simulation to match an optimal reference pattern. Multitalker Chinese vowel recognition was tested in NH subjects listening to a four-channel acoustic CI simulation, with and without the acoustic rescaling algorithm. Results showed a small but significant improvement in subjects' overall multitalker vowel recognition with the acoustic rescaling algorithm. Note that in the Luo and Fu study (2005), the largest improvements in performance were not always for the least-intelligible talkers. Nejime and Moore (1998) examined the effects of reduced speaking rates for speech intelligibility in noise, using a simulation of cochlear hearing loss in NH subjects. Reducing the speaking rate did not significantly improve intelligibility in the context of the simulated hearing loss. This lack of effect may have been due to the relatively weak contribution of speaking rate to intelligibility, or to processing artifacts associated with the signal modification.

In the present study, rather than normalize speech in one acoustic dimension or by linearly rescaling the formant space, we used a spectral normalization method based on statistical modeling of acoustic features to compensate for complex and dynamic acoustic variability at the speech segment level. As described earlier, no single acoustic feature can fully account for intelligibility difference across talkers. As a statistical modeling does not depend on any single feature, this approach may be more beneficial. The term "spectral normalization" is used because: (a) the spectral envelope was used to analyze the acoustic variability and (b) the proposed algorithm was intended to normalize spectral characteristics across different talkers. Note that spectral normalization in this study refers to a signal processing procedure rather than a perceptual process (e.g., "speaker normalization," as in Pisoni, 1993). The proposed spectral normalization algorithm was used for "front-end" processing (before CI speech processing), and was evaluated in CI users and NH subjects. In Experiment 1, recognition of sentences produced by two different talkers (one male and one female) was measured in CI listeners, with and without spectral normalization conditions; sentence recognition was measured in quiet for each talker independently. In Experiment 2, sentence recognition was measured in CI and NH listeners, with and without spectral normalization, using pitch-stretched transformations to simulate different talkers, i.e., speech was systematically pitch-stretched to produce different F0 and vocal tract configurations while preserving temporal characteristics such as speaking rate, overall duration, and amplitude. Subjective quality ratings and stimulus discriminability data were also collected from NH listeners.

II. METHOD

A. Spectral normalization algorithm

In the present study, the spectral normalization algorithm was based on a continuous statistical model to compensate for acoustic differences between talkers. By using a continuous model, speech from a variety of talkers may be adjusted to match a listener's optimal speech patterns. Specifically, the algorithm used a Gaussian mixture model (GMM) to represent spectral characteristics of a "source" talker at the segmental level, and transformed the source talker's spectral characteristics to that of a "target" talker using a trained spectral conversion function.

A GMM represents the distribution of the observed parameters $\mathbf{x}$ by m mixture Gaussian components in the form of

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x};\mu_i,\Sigma_i), \tag{1}$$

where α_i denotes the prior probability of component i ($\Sigma_{i=1}^{m}\alpha_i=1$ and $\alpha_i \geqslant 0$) and $N(\mathbf{x};\mu_i,\Sigma_i)$ denotes the normal distribution of the ith component with mean vector μ_i and covariance matrix Σ_i, in the form of

$$N(\mathbf{x};\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{p/2}\Sigma_i^{1/2}} \times \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right], \tag{2}$$

where p is the number of vector dimensions. The parameters of the model (α, μ, Σ) can be estimated using the well-known expectation maximization (EM) algorithm (Huang *et al.*, 2001).

After GMM modeling of the source talker's spectral distribution, the conversion function $F(x_t)$ is chosen such that the total conversion errors of n spectral vectors

$$\varepsilon = \sum_{t=1}^{n} (\mathbf{y}_t - F(\mathbf{x}_t))^2 \tag{3}$$

is minimized for training data, and where $\mathbf{x}_t$ is a spectral vector produced by the source talker and $\mathbf{y}_t$ is the time-aligned spectral vector produced by the target talker. Assuming that the source vector $\mathbf{x}_t$, follows a GMM model and that the source and target vectors are jointly Gaussian, the conversion function (Stylianou *et al.*, 1998; Kain and Macon, 1998; Mendel, 1995) is given in the form of

$$F(X_t) = \sum_{i=1}^{m} P(C_i|\mathbf{x}_t)[\mathbf{v}_i + \mathbf{T}_i \Sigma_i^{-1}(\mathbf{x}_t - \mu_i)], \tag{4}$$

where $P(C_i|\mathbf{x}_t)$ is the posterior probability of the ith Gaussian component given $\mathbf{x}_t$. $P(C_i|\mathbf{x}_t)$ is calculated by the application of Bayes theorem:

$$P(C_i|\mathbf{x}_t) = \frac{\alpha_i N(\mathbf{x}_t; \mu_i, \Sigma_i)}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}_t; \mu_j, \Sigma_j)}. \tag{5}$$

The unknown parameters $\mathbf{v}_i$ and $\mathbf{T}_i$ are computed by solving the following set of over-determined linear equations for all feature vector $t = (1, \ldots, n)$:

$$y_t = \sum_{i=1}^{m} P(C_i|\mathbf{x}_t)[\mathbf{v}_i + \mathbf{T}_i \Sigma_i^{-1}(\mathbf{x}_t - \mu_i)]. \tag{6}$$

Note that Eqs. (4) and (6) are identical on the right-hand side but are different on the left side [$F(\mathbf{x}_t)$ for Eq. (4), y_t for Eq. (6)]. Hence, the minimum mean square error (MMSE) solutions for $\mathbf{v}_i$ and $\mathbf{T}_i$ from Eq. (6) will guarantee that the total conversion error of Eq. (3) is minimized. Estimating parameters $\mathbf{v}_i$ and $\mathbf{T}_i$ from Eq. (6) determines the conversion function in Eq. (4). Given a certain source vector $\mathbf{x}_t$, the MMSE estimate of the target vector is equivalent to the right-hand side of Eq. (4).

B. Implementation of spectral normalization

Once the spectral conversion function had been estimated from training data, the spectral conversion was performed as depicted in Fig. 1.

In the above-mentioned system, a Mel-scaled line spectral frequency (LSF) feature (Huang *et al.*, 2001) was used to train the GMM, as it is perceptually based and has smooth interpolation characteristics (Kain and Macon, 1998). Specifically, the Mel-scaled LSF coefficients were obtained as follows. After frame-based speech analysis and linear predictive coding (LPC) coefficients extraction, the LPC spectrum was transformed to a Mel-warped spectrum (Huang *et al.*, 2001) according to the relationship $M(f) = 1125 \ln(1 + f/700)$, where f is frequency in hertz and $M(f)$ is the corresponding Mel frequency in Mels. The warped spectrum

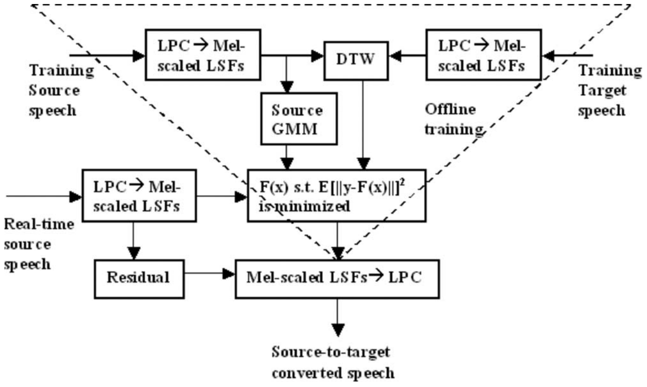

FIG. 1. Implementation framework of the GMM-based spectral normalization algorithm.

was then uniformly resampled using splined cubic phase interpolation to obtain the Mel-scaled LPC spectrum. A least-squares fit was used to transform the Mel-scaled LPC spectrum to Mel-scaled LPC coefficients, which were then transformed to Mel-scaled LSF coefficients.

To transform a given utterance, spectral feature vectors from the source talker's speech were extracted and transformed by the spectral conversion function that was trained using the GMM (as described in Sec. II A). The residual from the spectral extraction was then convolved with the modified spectral parameters to render the transformed speech signal. In the present study, there was no attempt to match the prosodic characteristics of source and target talkers. Hence, the source talker's average fundamental frequency (F0), speaking rate, and articulation rhythms were preserved after transformation.

To reduce computational load, a diagonal conversion was used [i.e., the $\mathbf{T}_i$ and Σ_i in Eq. (4) were in diagonal form]. This is a common practice in GMM training, as the correlation between distinct cepstral coefficients is very small (Stylianou *et al.*, 1998). The number of GMM components [i.e., m in Eq. (1)] was set to 64, as the contribution of additional GMM components to the acoustic distance between target and transformed speech is marginal beyond 64 components (Liu *et al.*, 2006).

C. Objective verification of the spectral normalization algorithm in cochlear implant simulations

Spectral conversion has been shown to effectively transform the spectral characteristics (e.g., formant position/bandwidth, spectral tilt, energy distribution, vocal tract length) of a source talker to that of a target talker without spectral degradation (Stylianou *et al.*, 1998; Kain and Macon, 1998; Liu *et al.*, 2006). For CI users, speech recognition performance is most strongly influenced by parameters that affect spectral resolution (e.g., the number of electrodes/channels). In general, speech recognition in quiet improves with increasing numbers of spectral channels (Fu, 1997; Dorman *et al.*, 1997b; Fishman *et al.*, 1997). To see the effect of spectral conversion on spectrally degraded speech (as is typically encountered by CI users), the proposed algorithm was evaluated using distance measurements with an acoustic CI simulation.

The acoustic CI simulation was implemented similarly to Shannon *et al.* (1995). The signal was first preemphasized with a filter coefficient of 0.95. The input frequency range (100–6000 Hz) was then bandpassed into 16, 8, 6, or 4 frequency analysis bands (24 dB/octave filter slope), distributed according to Greenwood's formula (Greenwood, 1990). The temporal envelope was then extracted from each frequency band by half-wave rectification and low-pass filtering (160 Hz envelope filter). The envelope of each band was then used to modulate a wideband noise, which was then spectrally limited by the same bandpass filter used for frequency analysis. Finally, the modulated carriers of each band were summed to render spectrally degraded speech. To measure the degree of spectral conversion for spectrally degraded speech, the acoustic distance between the transformed speech and the target speech was calculated as follows:

$$d^2_{\mathrm{MFCC}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{p}[c^{n,k}_{\mathrm{converted}} - c^{n,k}_{\mathrm{target}}]^2, \qquad (7)$$

where N is the total frame numbers in feature streams, $c^{n,k}$ is the kth component of the Mel-frequency cepstral coefficients (MFCC) vector in frame n, and p is the MFCC order (14, in the present case). Lower values of d^2 indicate greater spectral similarity between the transformed speech and the target speech. When the spectrum of the transformed speech perfectly aligns with that of the target speech, $d^2=0$.

The objective analysis was performed using IEEE sentences (IEEE, 1969), recorded with one male (M1) and one female (F1) talker. Spectral normalization from F1 to M1 was analyzed. The GMM training data set included 100 sentences randomly selected from the database; the testing data set included the entire database (720 sentences). The average acoustic distance between the source (F1) and target (M1) speech was calculated across the whole testing data set. The average acoustic distance for each condition was then converted to decibel units referenced to the acoustic distance between the unprocessed source and target speech (i.e., no acoustic CI simulation or spectral normalization). Figure 2 shows the mean acoustic distance (in decibels) between the source and target speech, as a function of the number of spectral channels. The acoustic distance decreased similarly with (dashed line) or without (solid line) spectral normalization as the number of spectral channels was reduced. The acoustic distance was significantly reduced (paired t-test: $p < 0.05$) with the spectral normalization algorithm; the mean reduction in acoustic distance was −2.73 dB, across all spectral resolution conditions. The objective analysis showed that spectral normalization was efficient in transforming the source speech in mimicking the target speech, regardless of the number of spectral channels.

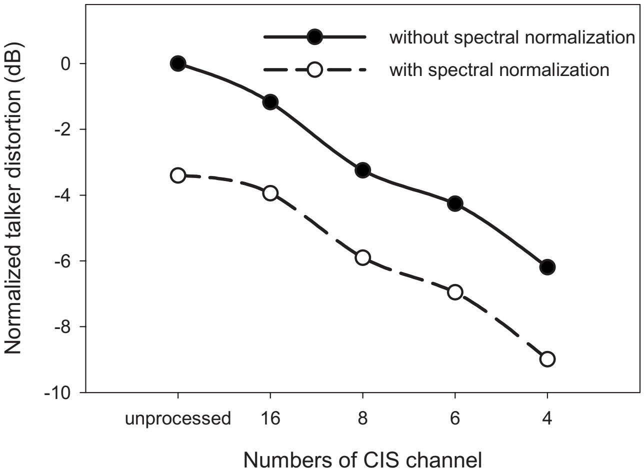

FIG. 2. Normalized talker distortion as a function of number of channels. Solid line: Without spectral normalization. Dashed line: With spectral normalization. Note that the talker distortion between talkers F1 and M1 (unprocessed speech) was used as the reference.

III. EXPERIMENT 1: EFFECT OF SPECTRAL NORMALIZATION ON SENTENCE RECOGNITION WITH TWO DIFFERENT TALKERS

A. Methods

1. Subjects

Nine postlingually deafened adult CI users (7 men, 2 women) participated in this experiment. Table I lists relevant demographics for the CI subjects. All subjects were native speakers of American English and had extensive experience in speech recognition experiments. All subjects provided informed consent and all were paid for their participation.

TABLE I. Subject demographics for the cochlear implant patients who participated in the present study.

Subject	Age	Gender	Etiology	Implant Type	Strategy	Duration of Implant use (years)
S1	67	M	Hereditary	Nucleus-22	SPEAK	14
S2	75	M	Noise induced	Nucleus-22	SPEAK	9
S3	72	F	Unknown	Nucleus-24	ACE	5
S4	54	M	Unknown	Nucleus-22	SPEAK	11
S5	62	F	Genetic	Nucleus-24	ACE	2
S6	55	M	Hereditary	Freedom	ACE	1
S7	52	M	Unknown	Clarion-CII	HiRes	6
S8	48	M	Trauma	Nucleus-22	SPEAK	13
S9	64	M	Trauma/unknown	Nucleus-22	SPEAK	15

2. Stimuli and speech processing

IEEE sentences (IEEE, 1969), recorded with one male (M1) and one female (F1) talker, were used in this experiment. The mean F0, across all sentences was 92 Hz for M1 and 185 Hz for F1. It is assumed that, in practice, spectral normalization is beneficial only when a less intelligible talker is transformed toward a more intelligible talker. Because it was unknown which talker might produce better recognition performance in individual CI subjects, the spectral transformation was performed between both talkers (i.e., M1 was transformed to F1, and F1 was transformed to M1). The GMM for the source talker was trained with 100 randomly selected sentences, resulting in over 60,000 Mel-scaled LSF feature vectors (25th order). The function to transform the source talker to the target talker was estimated according to Eq. (4). After training the conversion function, all sentences with each source talker were spectrally transformed toward the target talkers. Note that the training sentences were also transformed and included in the listening test to increase the available speech materials for the experiment. For descriptive purposes, when the source talker was M1 and the target talker was F1, the transformed speech was labeled M1-to-F1 (and vice versa). In Experiment 1, IEEE sentence recognition was tested for four talker conditions: M1 (unprocessed), F1 (unprocessed), M1-to-F1, and F1-to-M1.

3. Procedure

Subjects were tested using their clinically assigned speech processors and self-adjusted comfortable volume/sensitivity settings; once testing began, these settings were not changed. Subjects were tested while seated in a double-walled sound-treated booth (IAC). Stimuli were presented via a single loud speaker at 65 dBA. The sentences in the IEEE database were divided into 72 lists, with 10 sentences per list. For each run, a list was randomly chosen (without replacement) and the sentences from within the list were presented in random order. Subjects were asked to repeat what they heard; the experimenter tabulated all correctly identified words. Performance was calculated according to the ratio between correctly identified words and all words presented in the list; performance was typically averaged across four to five lists for each talker condition. To familiarize subjects with the different talkers and test procedures, a practice session with one randomly selected list (without replacement) was provided prior to the sentence recognition test in each condition. Note that the speech stimuli used in the practice session were not included in test stimulus set. The test order for the different talker conditions was randomized for each subject. No feedback was provided during the test.

B. Results and discussion

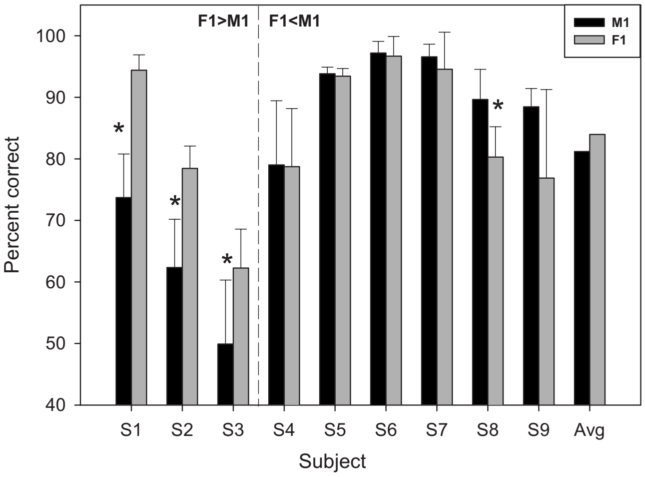

FIG. 3. Individual and mean sentence recognition performance for talkers M1 and F1. For subjects S1–S3, performance with F1 was better than that with M1; for subjects S4–S9, performance was better with M1 than with F1. The error bars show 1 s.d., and the asterisks show significantly different performance between the two talkers ($p < 0.05$).

Figure 3 shows individual subjects' sentence recognition performance for the M1 and F1 unprocessed source talkers, as well as the mean performance across subjects. Subjects are ordered according to talker sensitivity. Note that throughout the paper, "talker sensitivity" refers to the magnitude of the difference in performance between the two talkers. Mean performance with talker F1 was 2.8 percentage points greater than that with M1; however, the difference was not significant [one-way repeated measures (RM) analysis of variance (ANOVA): $F(1,8)=0.560$, $p=0.476$]. For subjects S1–S3, performance was better with F1 than with M1; for subjects S4–S9, performance was better with M1 than with F1. The difference in performance between talkers was significant for subjects S1, S2, S3, and S8 (t-test: $p < 0.05$; analysis performed within individual subjects using raw data from multiple sentence lists). There was intersubject variability in terms of talker sensitivity, ranging from 0.3 percentage points for subject S4 to 21 percentage points for subject S1.

Because the talker that produced better speech understanding differed among individual subjects, the most relevant comparison between unprocessed and spectrally transformed speech is in terms of the "Better" and "Worse" talker. Ideally, after spectral transformation, performance with the Worse talker would be equivalent to that with the Better talker (and vice versa). For convenience, the term "Worse-to-Better" refers to the transformation of the Worse toward the Better talker (and vice versa). Table II compares performance with unprocessed speech to that with spectrally normalized speech in terms of the Better and Worse talkers. Note that individual subject data were analyzed with t-tests, using raw performance data from multiple sentence lists; mean performance data (across subjects) were analyzed using mean data from each subject (across sentence lists). While the Better talker differed among individual subjects, the mean baseline performance difference between the Better and Worse talker was 8.1 percentage points; this difference was significant [one-way RM ANOVA: $F(1,8)=10.164$, $p=0.013$]. For the Better-to-Worse transformation, mean performance was significantly poorer than that with the Better talker [one-way RM ANOVA: $F(1,8)=5.558$, $p=0.046$]. The difference was significant for subjects S1, S2, and S4. There was no significant difference in mean performance between the Better-to-Worse transformation and the Worse talker

TABLE II. Performance difference between unprocessed source talkers (i.e., M1 vs F1), and between spectrally normalized and unprocessed talkers. Note that because the performance with talkers M1 and F1 differed among individual subjects, comparisons are made in terms of the "Better" and "Worse" talker. Bold numbers indicate significant differences in performance across different sentence lists ($p < 0.05$).

			Performance difference (percentage points)				
			Better	Better-to-Worse		Worse-to-Better	
Subject	Better talker	Worse talker	vs Worse	vs Better	vs Worse	vs Better	vs Worse
S1	F1	M1	**20.7**	**−10.2**	**10.5**	**−15.5**	5.2
S2	F1	M1	**16.1**	**−11.6**	4.5	−7.8	8.3
S3	F1	M1	**12.4**	5.3	**17.7**	−10.6	1.7
S4	M1	F1	0.3	**−21.6**	**−21.3**	−0.5	−0.2
S5	M1	F1	0.4	−0.7	−0.3	1.7	2.1
S6	M1	F1	0.5	−0.2	0.3	0.8	1.3
S7	M1	F1	2.0	−3.2	−1.2	−0.4	1.6
S8	M1	F1	**9.4**	−4.9	4.4	−6.0	3.3
S9	M1	F1	11.6	−8.7	2.9	−6.8	4.8
Avg across all nine subjects			**8.1**	**−6.2**	2.0	**−5.0**	**3.1**
Avg across subjects S1, S2, S3 and S8			**14.7**	−5.4	9.3	**−10.0**	**4.6**

[one-way RM ANOVA: $F(1,8)=0.308$, $p=0.594$]; however, performance differed significantly for subjects S1, S3, and S4. For the Worse-to-Better transformation, mean performance remained significantly poorer than that with the Better talker [one-way RM ANOVA: $F(1,8)=6.624$, $p=0.033$]; the difference was significant only for subject S1. Interestingly, performance with the Worse-to-Better transformation was significantly better than that with the Worse talker [one-way RM ANOVA: $F(1,8)=12.967$, $p=0.007$], although the difference was not significant for any individual subject.

Figure 3 shows that only four out of the nine subjects exhibited significant differences in intelligibility between the F1 and M1 talkers. This may have been due to ceiling performance effects in some subjects (S5, S6, and S7). Further analysis was performed using only the subjects whose baseline performance was significantly affected by talker (S1, S2, S3, and S8). Results are shown in Table II alongside the analyses with all nine subjects. For the Better-to-Worse transformation, mean performance was 5.4 percentage points lower than that with the Better talker; however, this difference was not significant [one-way RM ANOVA: $F(1,3)=1.957$, $p=0.256$]. Mean performance for the Better-to-Worse transformation was 9.3 percentage points better than that with the Worse talker; however, this difference was not significant [one-way RM ANOVA: $F(1,3)=8.757$, $p=0.060$]. For the Worse-to-Better transformation, mean performance remained 10.0 percentage points poorer than that with the Better talker; this difference was significant [one-way RM ANOVA: $F(1,3)=23.383$, $p=0.017$]. Mean performance for the Worse-to-Better transformation was 4.6 percentage points better than that with the Worse talker [one-way RM ANOVA: $F(1,3)=10.672$, $p=0.047$], Thus, the subanalyses using subjects S1, S2, S3, and S8 showed similar performance patterns to those with the previous analyses using all nine subjects.

In general, the spectral normalization algorithm produced the intended results, i.e., performance improved when a Worse talker was transformed toward a Better talker and vice versa. For some subjects, performance with the transformed talkers did not always follow this general trend. For example, for subject S3, performance improved when the Better talker was transformed to the Worse talker. Conversely, performance slightly declined for subject S4 when the Worse talker was transformed to the Better talker. This adverse effect may have been because the spectral normalization could not completely compensate for all the acoustic/perceptual differences between the source and target talkers. Alternatively, the transformation may have resulted in "talkers" that were not included or sampled in the test materials.

Sensitivity to the spectral normalization algorithm also varied among subjects. For example, there was only a 2 percentage point difference in performance among the four conditions for subjects S5 and S6. For subject S4, performance with the Better-to-Worse transformation was 22 percentage points poorer than that with the Better talker; there was only ~1 percentage point difference in performance among the remaining three talker conditions. In general, the effect of spectral normalization was strongest for subjects whose performance differed substantially between M1 and F1 (i.e., subjects S1, S2, S3, S8, and S9). In terms of mean performance, note that the Better-to-Worse transformation resulted in a decrement of ~6 percentage points, while the Worse-to-Better transformation resulted in an improvement of ~3 percentage points. While this is a relatively small difference in terms of effect size, there are three possible explanations for this bidirectional imbalance. First, the mean performance deficit with the Better-to-Worse transformation may have been primarily due to the large drop in performance for subject S4. Second, artifacts associated with the spectral normalization algorithm (e.g., spectral discontinuities, unnaturalness) may have limited any improvements in performance with the Worse-to-Better transformation and may have contributed more strongly to "worsening" performance with the Better-to-Worse transformation. Third, ceiling effects may have limited the degree of improvement with spectral normalization, but not the degree of deterioration in performance.

The results of Experiment 1 demonstrated that spectral normalization, using a GMM model trained with relatively few stimuli, significantly improved speech understanding with less-intelligible talkers, especially for CI users whose speech performance was sensitive to different talkers. The objective acoustic measures using the CI simulation showed that spectral normalization was efficient in transforming the source speech toward the target speech, regardless of the number of spectral channels. Although some CI subjects were more sensitive than others to talker differences and the subsequent spectral normalization, the perceptual measures showed relatively small effects, on average. The modest effects may have been due to the small number and/or quality of the test talkers (1 male and 1 female) who may not have elicited sufficient baseline talker sensitivity effects. Also, ceiling performance effects associated with sentence recognition may have limited the effects of spectral normalization.

IV. EXPERIMENT 2: EFFECT OF SPECTRAL NORMALIZATION ON SENTENCE RECOGNITION WITH SIMULATED TALKERS

In Experiment 1, the speech materials differed between the two talkers not only in terms of spectral cues, but also in terms of temporal cues, even for the same sentences. Previous studies have shown that speech intelligibility may be influenced by speaking rate (Kurdziel *et al.*, 1976; Miller and Volaitis, 1989; Gordon-Salant and Fitzgibbons, 1997). Cross-talker temporal variability such as voice-onset-time may also affect speech recognition (Allen *et al.*, 2003). Temporal cue effects have been observed in NH listeners (Miller and Volaitis, 1989), elderly listeners (Kurdziel *et al.*, 1976), HI listeners (Gordon-Salant and Fitzgibbons, 1997; Kirk *et al.*, 1997), and CI users (Liu *et al.*, 2004). In Experiment 1, the spectral normalization algorithm was intended to modify only the spectral information. However, it is possible that modified spectral information may interact with temporal information (which was not modified) and thereby affect speech understanding. It would be preferable to test the spectral normalization algorithm using different talker speech materials that have been normalized in terms of temporal information. It is very difficult to constrain temporal cues (i.e., speaking rate, total duration, emphasis, etc.) across different talkers with naturally produced speech materials. Therefore, in Experiment 2, the different talker conditions were simulated by adjusting the voice pitch and vocal tract characteristics of a reference talker (F1).

A. Methods

1. Subjects

The same 9 CI subjects from Experiment 1 also participated in Experiment 2. Four NH subjects (2 men, 2 women) also participated in Experiment 2, and served as a control group. All NH subjects had sensitivity thresholds better than 15 dB hearing level for audiometric test frequencies from 250 to 8000 Hz; all were native speakers of American English. Informed consent from each subject was obtained for the study.

2. Stimuli and speech processing

In Experiment 2, talker differences were simulated by systematically altering the acoustic characteristics of a reference talker (F1), while preserving speaking rate (i.e., duration of utterances) and prosodic characteristics, in the form of relative changes in F0. The IEEE sentences produced by talker F1 from Experiment 1 were used as the reference. To simulate different talkers, sentences were altered by using the "pitch-stretch" processing feature in COOL EDIT PRO (Version 2.0; Syntrillium Software). The pitch-stretching algorithm changed the fundamental frequency of the original speech and hence the spectral envelope, mimicking different vocal tract configurations. Each sentence produced by talker F1 was processed using six different pitch-stretch ratios: 0.6, 0.8, 1.0, 1.2, 1.4, and 1.6. Higher ratios resulted in lower-pitched speech, and smaller ratios resulted in higher-pitched speech; when the ratio was equal to 1.0, there was no pitch shift (i.e., the original speech tokens from talker F1). For example, the average F0 for talker F1 across all sentences was 185.10 Hz; when pitch-stretched by a ratio of 1.6, the average F0 was 116.96 Hz (i.e., $185.10/1.6=115.69$ Hz), simulating a male voice. Note that because the reference talker F1 was female, the minimum ratio was 0.6 in this experiment, as lesser values produced overly high F0 values. Thus, ratios of 0.6, 0.8, and 1.0 were intended to simulate female talkers, while ratios of 1.2, 1.4, and 1.6 were intended to simulate male talkers. Note that as the pitch-stretching ratio deviated from 1.0, speech increasingly sounded less natural. For reference purposes, the transformations associated with the different pitch-stretching ratios are labeled T0.6, T0.8, T1.0 (unprocessed speech by talker F1), T1.2, T1.4, and T1.6.

After generating the different pitch-shift transformations, the spectral normalization algorithm was applied, with T1.0 as the target talker. The same 100 sentences used in Experiment 1 were used to train the GMM model, while the entire database (720 sentences) was used for testing. Spectral normalization was performed exactly as in Experiment 1. The stimuli in Experiment 2 included 11 sets of talkers: one source talker (T1.0), five pitch-shift transformations (T0.6, T0.8, T1.2, T1.4, and T1.6), and five spectral transformations (T0.6-to-T1.0, T0.8-to-T1.0, T1.2-to-T1.0, T1.4-to-T1.0, and T1.6-to-T1.0). For the spectral transformations, T0.6-to-T1.0 and T0.8-to-T1.0 represented female-to-female transformations, while T1.2-to-T1.0, T1.4-to-T1.0, and T1.6-to-T1.0 represented male-to-female transformations.

Table III shows the pitch and formant analysis for the pitch-shifted and spectrally transformed speech. Table III shows that voice pitch was well-scaled by the pitch-stretching operation, and was maintained by spectral transformation. While formant frequencies were not maintained or scaled by the pitch-stretching algorithm (relative to the source speech T1.0), spectral normalization largely restored formant frequencies to those of the source speech. The mean difference of all formant frequencies between the target

TABLE III. Pitch and formant analysis for the pitch-shift and spectral transformations in Experiment 2. The target F0 for the pitch-shift transformations was scaled according to the pitch-stretching ratio used for processing; the target F0 for the spectral transformation refers to the measured F0 values after pitch-stretching. The F0s and formant frequencies were measured with software WAVESURFER 1.8.5. The F0s were averaged across all IEEE sentences. The formant frequencies were estimated for the vowel /IY/ from the sentence "Glue the sheet to the dark blue background." Note that reference talker T1.0 (in bold) was F1 from Experiment 1.

Transformation	Condition	Target F0	Measured F0	Measured F1	Measured F2	Measured F3
Pitch-shift	T0.6	185/0.6 = 308	298	326	582	4040
	T0.8	185/0.8=231	228	422	2051	3046
	T1.0	**185/1.0=185**	**185**	**344**	**2440**	**2859**
	T1.2	185/1.2=154	155	290	2062	2405
	T1.4	185/1.4=132	133	248	1816	2232
	T1.6	185/1.6=116	117	214	1610	2267
Spectral	T0.6-to-T1.0	298	300	291	2497	2996
	T0.8-to-T1.0	228	230	290	2408	2871
	T1.2-to-T1.0	155	157	279	2366	2750
	T1.4-to-T1.0	133	135	274	2151	2744
	T1.6-to-T1.0	117	118	278	1750	2498

speech (i.e., T1.0) and the spectral transformation was only 118 Hz, with a standard deviation of 199 Hz.

Figure 4 shows example wave forms for the sentence "Glue the sheet to the dark blue background," produced by source talker T1.0 and two pitch-shift transformations (T0.6, T1.6); note that the duration and modulation depth is nearly identical across the wave forms. The top panel of Fig. 5 shows the spectral envelope for a speech segment within the vowel /IY/ from the word "sheet;" note the relative stretch in the spectral envelope for the two pitch-shift transformations. The bottom panel of Fig. 5 shows the spectral envelope for the same speech segment, as produced by two spectral trans-

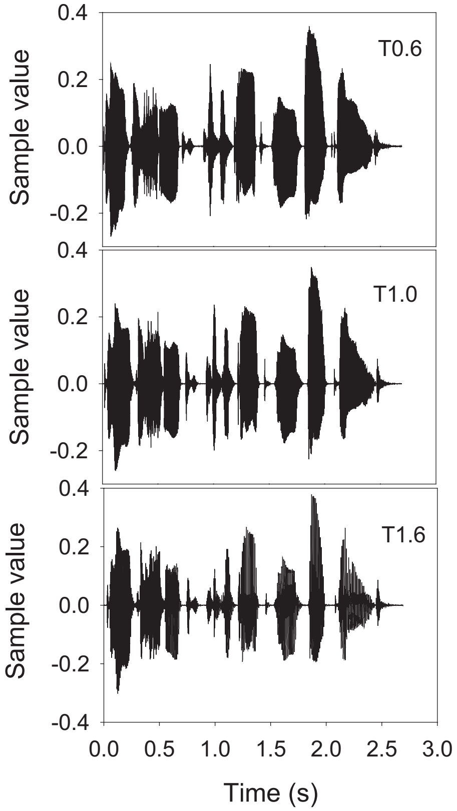

FIG. 4. Wave forms for the sentence "Glue the sheet to the dark blue background." Top panel: Pitch-shift transformation T0.6 (upward pitch shift). Middle panel: Reference talker T1.0 (unprocessed speech from talker F1). Bottom panel: Pitch-shift transformation T1.6 (downward pitch shift).

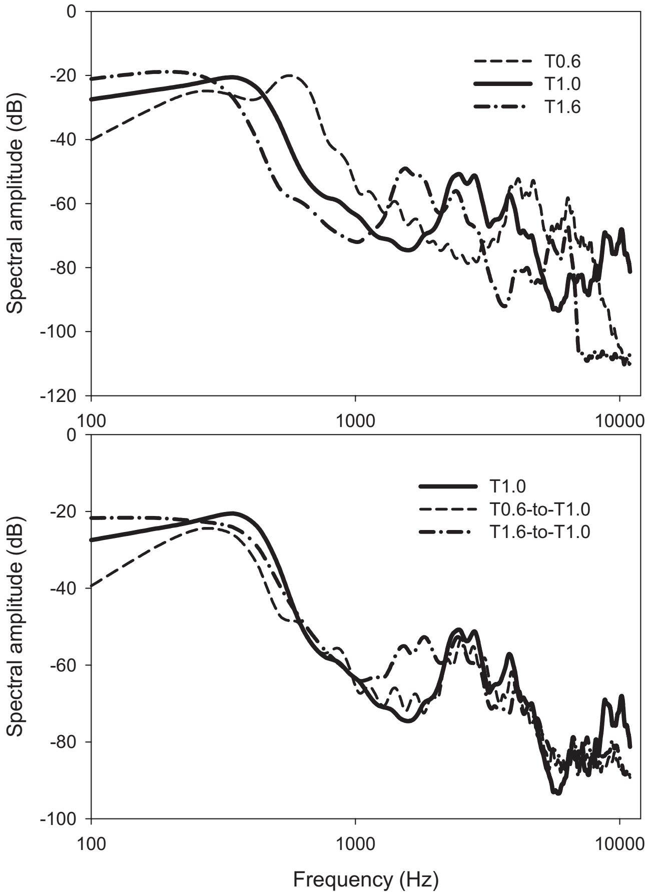

FIG. 5. Spectral envelopes for different processing conditions in Experiment 2. Top panel: Spectral envelopes for reference talker T1.0 and pitch-shift transformations T0.6 and T1.6. Bottom panel: Spectral envelopes for T1.0 and spectral transformations T0.6-to-T1.0 and T1.6-to-T1.0.

formations (T0.6-to-T1.0, T1.6-to-T1.0) and reference talker T1.0; note that the spectral envelopes are quite similar for the two spectral transformations and T1.0.

3. Procedure

For all talker conditions, IEEE sentence recognition was measured using the same procedures described in Experiment 1. After the speech recognition test, subjective quality ratings were obtained from the same NH subjects. Subjects were asked "How would you rate the overall speech quality on a scale from 1 to 10, with larger values indicating better overall speech quality?" Subjective ratings were anchored to T1.0, which was given the highest rating (10, on a 10-point scale). For each pitch-shift and spectral transformation, speech quality ratings were averaged across subjects. In addition to sentence recognition, discriminability among the pitch-shift transformations was also measured in NH subjects to verify whether the pitch-stretching algorithm produced different "talker identities." During the discriminability test, subjects were presented with a sentence produced by T1.0 and a different sentence produced by one of the pitch-shift transformations. NH subjects were asked whether the sentences were produced by the same or different talkers. Each of the five pitch-shift transformations was compared to talker T1.0 six times. The presentation order of the processed and unprocessed sentences was randomized, and the test sentences were randomly selected (without replacement) from the test materials.

B. Results and discussion

In terms of stimulus discriminability, for three out of the four NH subjects, all of the pitch-shift transformations sounded like different talkers, relative to the reference talker T1.0 (i.e., 100% discrimination across all six trials). For the remaining NH subject, pitch-shift transformations T1.2, T1.4, and T1.6 were easily discriminated from talker T1.0 (100% discrimination across all six trials); however, T0.8 and T0.6 were judged to be the same as T1.0 in four out of six trials. Thus, the pitch-shift transformations were judged, for the most part, to represent different talker identities.

Figure 6 shows the overall speech quality ratings for the 4 NH subjects, with and without spectral normalization, as a function of pitch-shift transformation. A two-way RM ANOVA showed that overall speech quality ratings were significantly affected by pitch-shift [$F(5,15)=15.352$, $p<0.001$]. While spectral normalization also seemed to affect quality ratings, the effect failed to reach significance [$F(1,15)=9.517$, $p=0.054$]. Except for pitch-shift transformation T1.2, all the pitch-shift and spectral transformations produced significantly lower ratings, relative to reference talker T1.0 (t-test: $p<0.05$). Quality ratings were generally lower after spectral transformation; the decrements were only significant for T1.2-to-T1.0 (t-test: $p<0.05$). This suggests that pitch shifting may have introduced one set of artifacts, and the spectral normalization (intended to compensate only for spectral differences) may have introduced a second set of artifacts. Artifacts associated with spectral normalization may have added to the pitch-shift artifacts, further reducing the speech quality ratings.

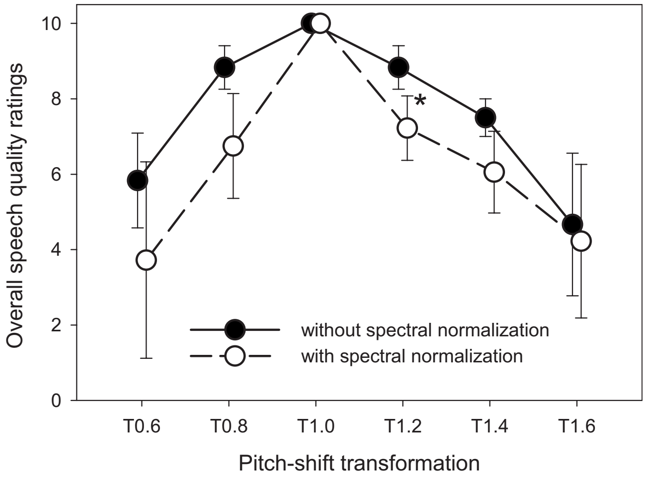

FIG. 6. NH subjects' overall speech quality ratings for the pitch-shift transformations, with (open symbols) and without (closed symbols) spectral normalization. The error bars show 1 s.d., and the asterisks indicate significantly different ratings with spectral normalization ($p<0.05$). Note that source talker T1.0 (unprocessed speech from talker F1) was used to anchor the subjective quality ratings.

Figure 7 shows sentence recognition performance for CI and NH listeners (circle and diamond symbols, respectively), with and without spectral normalization (open and closed symbols, respectively), as a function of pitch-shift transformation. For NH subjects, sentence recognition remained nearly perfect for all conditions, except for T0.6-to-T1.0 (94% correct, significantly lower than performance with T0.6). CI subjects were very sensitive to the different pitch-shift and spectral transformations. Mean peak performance was 84% correct with T1.0. Performance with the pitch-shift transformations sharply declined as the shift ratios became more extreme. With T0.6, mean performance was only 20% correct, and with T1.6, mean performance was only 31%

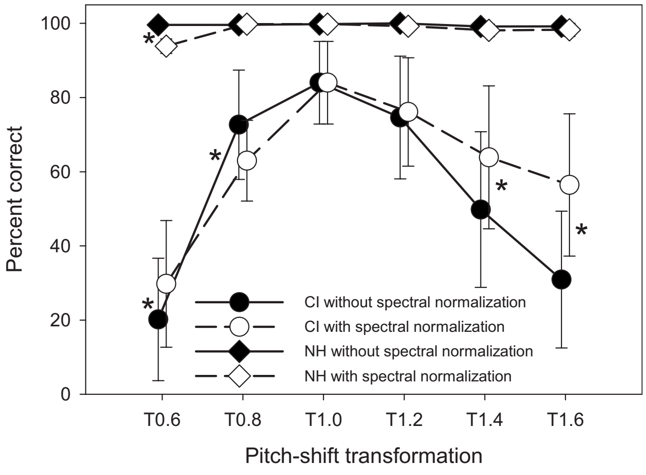

FIG. 7. Sentence recognition performance for NH and CI subjects, with (open symbols) and without (closed symbols) spectral transformation, as a function of pitch-shift transformations. The error bars show 1 s.d., and the asterisks indicate significantly different performance after spectral transformation ($p<0.05$).

correct. Performance with spectral normalization also declined for the more extreme pitch-shift transformations, but the decline was less steep than without spectral normalization. A two-way repeated measures ANOVA showed that performance was significantly affected by pitch-shift $[F(5,40)=124.014,\ p<0.0001]$ and spectral normalization $[F(1,40)=64.770,\ p<0.001]$. There was a significant interaction between the pitch-shift and spectral transformations $[F(5,40)=21.557,\ p<0.001]$. Post-hoc Bonferroni t-tests showed that performance with T1.0 was significantly better than that with T0.6, T0.8, T1.4, and T1.6, before or after spectral normalization ($p<0.05$). Post-hoc Bonferroni t-tests also showed that spectral normalization significantly improved performance for T0.6, T1.4, and T1.6 ($p<0.001$). In contrast, performance with T1.2 was not significantly different from that with T1.0, with ($p=0.271$) or without spectral normalization ($p=0.078$). Performance with T0.8 significantly declined after spectral normalization ($p=0.012$).The failure of spectral normalization to significantly improve performance with T0.8 and T1.2 may have been due to the relatively small differences in baseline performance between T0.8, T1.0, and T1.2, before normalization. The potential benefits of spectral normalization may not have been large enough to overcome processing artifacts associated with the speech modification and synthesis.

The results suggest that, despite potential signal processing artifacts, spectral normalization may benefit CI users. It is important to note that, as suggested from Fig. 6, the artifacts associated with spectral normalization may have added to the artifacts associated with the pitch-shift algorithm. Note that the spectral normalization was intended to modify the spectral envelope toward that of the reference talker, not to reduce the artifacts associated with the pitch-shift processing. Ultimately, spectral normalization significantly improved CI users' speech understanding with the pitch-shift transformations. This implies that CI listeners may not have been sensitive to these processing artifacts (due to the reduced spectral resolution), or that CI listeners were able to ignore these artifacts and receive the benefits of spectral normalization.

Given the probable artifacts associated with pitch-shift and spectral normalization, it is possible that some learning may have occurred during testing. Note that the test order was randomized within and across subjects for the measurements with pitch-shift and spectral transformations. A two-way ANOVA was conducted for each individual subject (with the pitch-shift ratio and test session as factors), for both the baseline and spectral normalization conditions. While there were significant effects for the pitch-shift ratio (both with and without spectral normalization), there were no significant effects for test session, for any subject in any condition ($p>0.05$). Note also that the effects of spectral normalization were measured acutely. It is possible that long-term experience or explicit training might have influenced baseline performance and/or further enhanced the benefit of spectral normalization.

GENERAL DISCUSSION

The results of the present study demonstrate that the proposed spectral normalization algorithm can significantly improve CI users' speech understanding with less-intelligible talkers. In Experiment 1, spectral normalization provided the greatest benefit to CI subjects who exhibited the greatest talker sensitivity. In Experiment 2, a pitch-shift algorithm was used to simulate different talkers while keeping temporal cues (e.g., speaking rate, overall sentence duration, temporal modulation depth) constant. While pitch-shifting and subsequent spectral normalization produced some undesirable processing artifacts, CI users' speech recognition improved with the spectral normalization algorithm. Taken together, the results suggest that this spectral normalization approach may benefit CI users' understanding of speech produced by less-intelligible talkers. However, some considerations should be kept in mind when interpreting these results, and in designing an effective spectral normalization algorithm for real-time speech processing.

In Experiment 1, only four of the nine CI subjects exhibited significant better performance with one of the two test talkers (S1, S2, and S3 with talker F1; S8 with talker M1). The best-performing subjects exhibited no significant difference in performance with talkers F1 or M1. It is possible that a greater number of source talkers would have elicited stronger talker sensitivity effects in all subjects, albeit with different talkers for each subject. It is also possible that interfering noise may have elicited more talker sensitivity across subjects. Note that both F1 and M1 produced the IEEE stimuli in the manner of "clear" speech, i.e., relatively slow speaking rate, well articulated, etc. Thus, the normalization algorithm largely addressed spectral envelope differences between talkers, which might have to be more extreme to produce talker sensitivity effects. Temporal differences between talkers (e.g., speaking rate, overall duration, emphasis, etc.) may produce equal if not greater talker sensitivity effects. An effective speaker normalization algorithm may also need to compensate for temporal differences between talkers, as well as spectral differences.

Experiment 2 was designed to factor out possible contributions of varying temporal information and to expand the range of talker characteristics presented in Experiment 1. While the pitch-shift algorithm may not have been the ideal method to create different talker characteristics, it is difficult to control temporal variations among different talkers. One would have to record a very large database to include the range of spectral and temporal characteristics encountered in everyday listening experience. Alternatively, judicious amounts of duration adjustments (via cut and splicing or duplication of speech segments) might offer some experimental control, albeit with another set of possible signal processing artifacts. In Experiment 2, the pitch-shift transformations significantly reduced performance relative to the reference source talker F1. The results were in agreement with those from Experiment 1, in that spectral normalization generally improved speech understanding with less-intelligible talkers. However, it should be noted that the pitch-shift algorithm simulated only some of the acoustic characteristics that may differ between real talkers. Also, NH subjects' overall speech quality ratings suggest that the spectral normalization algorithm may introduce undesirable artifacts when talker differences are sufficiently extreme. While CI listeners generally

TABLE IV. r^2 significance values for linear regressions performed between the unprocessed talkers from Experiment 1 (M1 and F1) and the pitch-shift transformations from Experiment 2 (T0.6, T0.8, T1.2, T1.4, T1.6).

Talker (Exp. 1)	Pitch-shift transformation (Exp. 2)				
	T0.6	T0.8	T1.2	T1.4	T1.6
M1	r^2=0.409	r^2=0.532	r^2=0.718	r^2=0.521	r^2=0.635
	p=0.064	**p=0.026**	**p=0.004**	**p=0.028**	**p=0.010**
F1	r^2=0.393	r^2=0.725	r^2=0.688	r^2=0.470	r^2=0.358
	p=0.071	**p=0.004**	**p=0.006**	**p=0.041**	p=0.089

benefited from spectral normalization in Experiment 2, it is unclear whether the proposed spectral normalization algorithm would sufficiently compensate for differences between real talkers. Further testing with a more diverse group of real source talkers is needed to verify the feasibility of the proposed technique.

Individual CI subjects' talker sensitivity may have also contributed to the pattern of results in Experiments 1 and 2. For example, it might be expected that subjects who performed better with talker F1 in Experiment 1 would also perform better with the upwardly shifted transformations T0.6 and T0.8 in Experiment 2, as these pitch shifts were smaller relative to F1 than to M1. It might also be expected that these same subjects would benefit more greatly from the spectral transformations T1.2-to-T1.0, T1.4-to-T1.0 and T1.8-to-T1.0. Data were compared between experiments to see whether individual subjects' performance in Experiment 1 was reflected in Experiment 2. In the first analysis, subjects were divided into two groups: Group 1 (S1, S2, and S3; better performance with F1) and Group 2 (S4–S9; better performance with M1). A two-way ANOVA, with subject group (Group 1 or 2) and pitch-shift transformation (T0.6 or T0.8) showed no significant effect for subject group [$F(1,7)$ =0.0584, p=0.816]; post-hoc Bonferroni t-tests showed no significant effect for subject group, for either T0.6 (p =0.987) or T0.8 (p=0.643). One issue with this analysis is that for five out of the six subjects in Group 2, there was no significant difference in performance between M1 and F1. Individual subject performance with talker M1 or F1 in Experiment 1 was also compared to that with the pitch-shift transformations in Experiment 2. Table IV shows the r^2 and significance values for the linear regression analysis across different subjects. Subject performance with the F1 talker in Experiment 1 was fairly well correlated with performance for T0.8 and T1.2 from Experiment 2 (both were relatively close to the original F1 talker). Similarly, subject performance with the M1 talker in Experiment 1 was fairly well correlated with performance with T1.6 from Experiment 2 (the most "male" of the pitch-shift transformations). However, one issue with this analysis is that the better performers in Experiment 1 performed equally well with the M1 and F1 source talkers. Thus, it is difficult to separate talker sensitivity from overall performance with this regression analysis. In the present study, it is difficult to know how talker sensitivity for the top-performing subjects may have been limited by performance ceiling effects. Again, sufficiently different talkers or difficult listening conditions might allow talker sensitivity effects to emerge in even good CI users. It may be true that better CI performers may be less sensitive to talker differences, and therefore may benefit less from spectral normalization. For these CI users, the acoustic input may be better matched to the electrode locations in the cochlea, or other patient-related factors may contribute to the better overall performance.

While the results from these two experiments are promising, special care is needed when designing a real-time normalization algorithm that will be robust to ambient noise, interfering speech, and the wide variety of talker characteristic found in everyday listening environments. A standard set of talkers and listening conditions might help to quickly identify a reference talker (or maybe even several reference talkers) that could be used in the algorithm. Ideally, the algorithm would be continuously updated as new talkers and listening conditions are introduced. Finally, CI patients would likely experience a period of adaptation to such a normalization algorithm (and any adverse processing artifacts). In the present study, the effects of spectral normalization were acutely measured, which may have underestimated the benefits after long-term experience.

V. CONCLUSIONS

The present study showed substantial differences in cross-talker intelligibility in CI users' speech recognition. A spectral normalization algorithm was used to compensate for acoustic differences between the less-intelligible and more-intelligible speech patterns from different talkers. In Experiment 1, spectral normalization was shown to significantly improve overall CI speech performance; however, some CI users were more sensitive than others to talker differences and the subsequent spectral normalization. In Experiment 2, the spectral normalization algorithm was applied to simulated talkers, in which the fundamental frequency and vocal tract characteristics were modified while preserving temporal information such as speaking rate. Compared to NH listeners, CI users' speech understanding was more sensitive to the pitch-shift transformations and subsequent spectral normalization. The results suggest that spectral normalization, as a front end to CI speech processing, may help CI users maintain perceptual constancy when presented with multiple and/or less-intelligible talkers.

ACKNOWLEDGMENTS

We are grateful to all subjects for their participation in these experiments. We thank three anonymous reviewers for their insightful comments and recommendations. Research was supported in part by NIH (No. DC-004993).

Allen, J. S., Miller, J. L., and DeSteno, D. (**2003**). "Individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **113**, 544–552.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**, 975–989.

Bond, Z. S., and Moore, T. J. (**1994**). "A note on the acoustic-phonetic characteristics of inadvertently clear speech," Speech Commun. **14**, 325–337.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech 1. Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Cox, R. M., Alexander, G. C., and Gilmore, C. (**1987**). "Intelligibility of average talkers in typical listening environments," J. Acoust. Soc. Am. **81**, 1598–1608.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997a**). "Stimulating the effect of cochlear implant electrode insertion depth on speech understanding," J. Acoust. Soc. Am. **102**, 2993–2996.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997b**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Fu, Q. -J. (**1997**). "Speech perception in acoustic and electric hearing," Ph.D. dissertation, University of Southern California, Los Angeles, CA..

Fu, Q. -J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**, 1889–1900.

Fishman, K., Shannon, R. V., and Slattery, W. H. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," Hear. Res. **40**, 1201–1215.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners," J. Speech Lang. Hear. Res. **40**, 423–431.

Green, T., Katiri, S., Faulkner, A., and Rosen, S. (**2007**). "Talker intelligibility differences in cochlear implant listeners," J. Acoust. Soc. Am. **121**, EL223–EL229.

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

Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.

Hood, J. D., and Poole, J. P. (**1980**). "Influence of the speaker and other factors affecting speech intelligibility," Audiology **19**, 434–455.

Huang, X. -D., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing—A Guide to Theory, Algorithm, and System Development*, (Prentice Hall, Englewood Cliffs, NJ).

IEEE (**1969**). *IEEE Recommended Practice for Speech Quality Measurements* (IEEE, New York).

Kain, A., and Macon, M. W. (**1998**). "Spectral voice conversion for text-to-speech synthesis," IEEE, ICASSP, **1**, 285–288.

Kirk, K. I., Pisoni, D. B., and Miyamoto, R. C. (**1997**). "Effects of stimulus variability on speech perception in listeners with hearing impairment," J. Speech Lang. Hear. Res. **40**, 1395–1405.

Kurdziel, S., Noffsinger, D., and Olsen, W. (**1976**). "Performance by cortical lesion patients on 40 and 60% time-compressed materials," J. Am. Aud Soc. **2**, 3–7.

Liu, C., Fu, Q. -J., and Narayanan, S. S. (**2006**). "Smooth GMM based multi-talker spectral conversion for spectrally degraded speech," IEEE ICASSP **5**, 141-144.

Liu, S., Rio, E. D., Bradlow, A. R., and Zeng, F.-G. (**2004**). "Clear speech perception in acoustic and electric hearing," J. Acoust. Soc. Am. **116**, 2374–2383.

Luo, X., and Fu, Q. -J. (**2005**). "Speaker normalization for Chinese vowel recognition in cochlear implants," IEEE Trans. Biomed. Eng. **52**, 1358–1361.

Mendel, J. M. (**1995**). *Lessons on Estimation Theory for Signal Processing, Communications and Control* (Prentice Hall, Englewood Cliffs, NJ).

Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**, 505–512.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.

Nejime, Y., and Moore, B. C. (**1998**). "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," J. Acoust. Soc. Am. **103**, 572–576.

Pisoni, D. B. (**1993**). "Long term memory in speech perception: Some new findings on talker variability, speaking rate, and perceptual learning," Speech Commun. **13**, 109–125.

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

Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1994**). "Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude," J. Acoust. Soc. Am. **96**, 1314–1324.

Stylianou, Y., Cappe, O., and Moulines, E. (**1998**). "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech Audio Process. **6**, 131–142.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **60**, 198–212.