Talker-identification training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition^{a)}

Vidya Krull^{b),c)} and Xin Luo

Department of Speech, Language, and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, Indiana, 47907

Karen Iler Kirk

Departments of Communication Sciences and Disorders and Otolaryngology-Head and Neck Surgery, Wendell Johnson Speech and Hearing Center, The University of Iowa, Iowa City, Iowa, 52242

(Received 6 June 2011; revised 23 January 2012; accepted 3 February 2012)

Understanding speech in background noise, talker identification, and vocal emotion recognition are challenging for cochlear implant (CI) users due to poor spectral resolution and limited pitch cues with the CI. Recent studies have shown that bimodal CI users, that is, those CI users who wear a hearing aid (HA) in their non-implanted ear, receive benefit for understanding speech both in quiet and in noise. This study compared the efficacy of talker-identification training in two groups of young normal-hearing adults, listening to either acoustic simulations of unilateral CI or bimodal (CI+HA) hearing. Training resulted in improved identification of talkers for both groups with better overall performance for simulated bimodal hearing. Generalization of learning to sentence and emotion recognition also was assessed in both subject groups. Sentence recognition in quiet and in noise improved for both groups, no matter if the talkers had been heard during training or not. Generalization to improvements in emotion recognition for two unfamiliar talkers also was noted for both groups with the simulated bimodal-hearing group showing better overall emotion-recognition performance. Improvements in sentence recognition were retained a month after training in both groups. These results have potential implications for aural rehabilitation of conventional and bimodal CI users. © 2012 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.3688533>]

PACS number(s): 43.71.Bp, 43.71.Ky, 43.66.Ts, 43.71.Sy [TD]

Pages: 3069–3078

I. INTRODUCTION

Understanding speech in the presence of competing talkers and extracting supra-segmental speech information, such as voice gender, talker identity, vocal emotion, and intonation continues to be challenging (Stickney *et al.*, 2004; Vongphoe and Zeng, 2005; Luo *et al.*, 2007) for average cochlear implant (CI) users, who have poor pitch perception. The spectral resolution in CI, limited by the electrode number and neural survival, is not high enough to resolve the fundamental frequency (F0) and its harmonics. Although the CI encodes the temporal envelope of the acoustic signal and temporal correlates of voice pitch are preserved at least in the low-frequency channels (Green *et al.*, 2002), the temporal-envelope pitch is relatively weak and insufficient for challenging listening tasks (Oxenham, 2008).

In listeners with normal hearing, voice pitch, the perceptual correlate of the F0, aids in the segregation of com-

peting talkers (Brox and Nooteboom, 1982; Assmann and Summerfield, 1990). CI users' limited access to voice pitch could contribute to their poor performance in tasks that require sound source segregation, such as when listening to a talker in the presence of competing talkers (e.g., Stickney *et al.*, 2004). Identification of the target talker has been shown to be poor with degraded spectral resolution and pitch cues even in quiet (e.g., Vongphoe and Zeng, 2005), let alone when there are competing talkers. Acoustic correlates of voice timbre including the range of F0, and the shape of laryngeal spectrum contribute to voice discrimination (Remez *et al.*, 1997). Other factors such as the nasal spectrum of the talker, hoarseness and breathiness of the voice, preciseness of articulation, and speaking rate also provide cues for talker identification (Kreiman and Sidtis, 2011).

Dynamic changes in pitch across an utterance are the primary indicators of intonation and play a critical role in conveying talkers' emotional states (Murray and Arnott, 1993). In the absence of salient pitch cues, listeners may misinterpret emotional information embedded in the speech signal, resulting in a breakdown in communication. Adult CI users have been shown to have difficulty in recognizing vocal emotions (Luo *et al.*, 2007). Poor perception of supra-segmental features of speech, including intonation, also has been reported in pediatric hearing aid (HA) and CI users (Most and Peled, 2007). However, HA users in the Most and

^{a)}Portions of this research were presented at the 33rd Midwinter Meeting of the Association for Research in Otolaryngology in February, 2010 and at the annual meeting of the American Auditory Society, Scottsdale, AZ, March 2011.

^{b)}Author to whom correspondence should be addressed. Electronic mail: vknull@indiana.edu

^{c)}Current address: Department of Speech and Hearing Sciences, Indiana University, 200 South Jordan Avenue, Bloomington, IN 47405.

Peled study showed an advantage for intonation perception compared to CI users, presumably due to better transmission of F0 and speech envelope cues by the HA.

CI users with low-frequency residual hearing may improve their access to voice pitch cues by using a HA in conjunction with their CI in the same ear [electric-acoustic stimulation or EAS (von Ilberg *et al.*, 1999; Turner *et al.*, 2004)], or in opposite ears [bimodal fitting (Kong *et al.*, 2005; Dorman *et al.*, 2008)]. In both real CI users and acoustic CI simulations, speech recognition (especially in a competing-talker background) has been found to improve with either EAS or bimodal fitting, even when the residual hearing was limited below 500 Hz (Turner *et al.*, 2004; Kong *et al.*, 2005; Dorman *et al.*, 2005; Brown and Bacon, 2009a,b; Buchner *et al.*, 2009).

It is unclear whether access to low-frequency residual hearing in EAS users and bimodal listeners also enhances the recognition of supra-segmental speech information. One line of indirect evidence comes from the better melody recognition with bimodal fitting than with CI alone (Kong *et al.*, 2005). Straatman *et al.* (2010) have shown improved perception of intonation in pediatric bimodal listeners when tested in the bimodal condition compared to CI alone. In contrast, Dorman *et al.* (2008) found similar voice-discrimination performance (within- or between-gender) in the bimodal and CI-alone conditions in a group of adult bimodal listeners. Ceiling effects may have limited the bimodal benefit for between-gender voice discrimination, whereas the frequency resolution in residual acoustic hearing may not be enough to improve within-gender voice discrimination. Overall, performance in these tasks remains poorer for EAS and bimodal listeners than for normal-hearing listeners.

In addition to improving information coding with EAS or bimodal fitting, researchers have used perceptual training to enhance listening abilities in conventional CI users (Fu *et al.*, 2005; Fu and Galvin, 2007). These auditory-training programs have mostly focused on improving the recognition of segmental speech information for CI users. However, less effort has been made to develop specific programs that enhance the use of supra-segmental speech information, such as talker identification. There is evidence that auditory training can improve the recognition of talkers' voices in listeners with normal hearing (e.g., Nygaard *et al.*, 1994). At the end of talker-identification training in that study, listeners also were significantly better at recognizing novel words produced by talkers they had heard during training but not novel words produced by unfamiliar talkers. In another study, Nygaard and Pisoni (1998) trained two groups of normal-hearing listeners (the experimental and control groups), each with a different talker set. At the end of talker-identification training, both groups were given a sentence-recognition test using only talkers from the training set of the experimental group. Thus the experimental group was tested with familiar talkers and the control group was tested with unfamiliar talkers. It turned out that the experimental group demonstrated superior sentence-recognition performance. These studies suggest that familiarity with a talker's voice may enhance the ability to understand the talker's

speech, referred to by Nygaard and colleagues as the "talker familiarity effect."

A few studies have examined the effects of talker-identification training in adult CI users (Barker, 2006) and using a CI simulation (Loebach *et al.*, 2008). Both studies showed that training in talker identification also improved sentence recognition. In Loebach *et al.* (2008), only familiar talkers were used in the post-training sentence-recognition test, so the effect of talker familiarity could not be assessed. Barker (2006) did compare the recognition of sentences produced by familiar and unfamiliar talkers but found no effect of talker familiarity. This is in contrast with previous results found using unprocessed speech in normal-hearing listeners (Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998). The different results may be due in part to the limited voice information available via a CI or a CI simulation.

The efficacy of talker-identification training in improving talker identification and/or speech recognition has not been assessed in EAS users or in bimodal listeners, both relatively new clinical populations. The EAS or bimodal benefits to speech recognition in noise have been attributed to better sound source segregation using the F0 cues (e.g., Turner *et al.*, 2004; Kong *et al.*, 2005), better glimpsing of the target speech using the amplitude envelope and voicing cues, or better perception of the first formant and its transition cues (e.g., Kong and Carlyon, 2007; Brown and Bacon, 2009a). For bimodal listeners, who have better access to these low-frequency acoustic cues, it is expected that talker-identification training and generalization to speech recognition would be more effective than it would be for conventional CI users.

This study addressed the following questions: (1) Does the addition of simulated low-frequency acoustic cues to CI simulation enhance the perceptual learning of voices in a structured auditory training program? (2) If talker identification is improved, does the training also lead to improved speech recognition for familiar and/or unfamiliar talkers? (3) Finally, given that emotion recognition and talker identification both rely strongly on pitch perception, will talker-identification training also benefit emotion recognition, especially in simulations of bimodal hearing? Similar to previous studies (Dorman *et al.*, 2005; Kong and Carlyon, 2007; Loebach *et al.*, 2008; Brown and Bacon, 2009a), an acoustic simulation approach was used in this study. Results from carefully designed simulation studies typically mirror performance in actual CI users (Turner *et al.*, 2004; Dorman *et al.*, 2005). Thus such an approach is useful as a first step to study the possible benefits of auditory training and low-frequency acoustic cues to CI users' talker identification and speech recognition without worrying about the confounding factors such as frequency-place mismatch, neural survival, and degree of residual hearing in CI users.

II. METHODS

A. Subjects

Thirty native speakers of American English (aged 18–25 yr) with normal hearing in both ears (thresholds ≤ 20 dB HL from 250 to 8000 Hz) participated in the study.

Twenty-four experimental subjects completed the 4-day talker-identification training paradigm. They were randomly assigned to be trained on one of two talker sets; this randomization was constrained to result in an equal number ($n = 12$) of subjects for each talker set. Further, for each talker set, half of the subjects were assigned to the “CI+HA” group and were presented with a CI simulation in their right ear and a HA simulation in their left ear. The other half were assigned to the “CI” group and were presented with only the CI simulation in their right ear. A third control group ($n = 6$) was tested with the CI+HA simulation without talker-identification training. Such a “test only” control group has been used in other training studies with CI simulation (Fu *et al.*, 2005), and normal-hearing listeners (Nygaard and Pisoni, 1998), to rule out any changes in sentence-recognition performance that could be attributed to test-retest.

B. Stimuli

1. Lexically controlled sentences

Sentences from the Department of Veterans Affairs Sentence Test (VAST) (Bell and Wilson, 2001) were used for talker-identification training and sentence recognition testing. Key words in each sentence are controlled for word frequency (i.e., how often the word occurs in English) and neighborhood density (i.e., the number of phonemically similar words). Key words belong to one of four lexical categories representing orthogonal combinations of word frequency (high or low) and neighborhood density (sparse or dense). Audio recordings of 320 sentences (80 sentences $\times$ 4 lexical categories) each produced by five female and five male talkers (Table I) were available.

To ensure that the effects of training would not be confounded by differences in intelligibility between familiar and unfamiliar talkers, two talker sets of equivalent intelligibility were created in a pilot study. Using a Latin square design,

10 normal-hearing adults were presented with different sentences (processed with an 8-channel CI simulation) spoken by each of the 10 talkers. Average intelligibility for each talker as well as across all talkers was calculated. Talkers with the poorest score within their gender and the largest deviation from their gender mean (T1 and T10) were removed. The remaining talkers were further divided into two talker sets (A and B) such that each talker set had nearly equal mean intelligibility ($\sim 74\%$ correct) and was balanced by talker gender.

2. Emotional sentences

The House Ear Institute emotion database consists of recordings from one male and one female talker, each producing 50 semantically neutral sentences to convey five target emotions (angry, happy, sad, anxious, and a neutral emotional state). A subset of 10 sentences from this database, shown to yield the highest scores for vocal emotion recognition in a group of normal-hearing listeners (Luo *et al.*, 2007), was used for emotion-recognition testing. This resulted in a total of 100 tokens (2 talkers $\times$ 5 emotions $\times$ 10 sentences).

3. Competing talker babble

Background noise was used in all of the talker-identification training and testing sessions and in half of the sentence-recognition testing sessions. Recordings of IEEE sentences (IEEE, 1969) produced by a male and a female talker (mean F0 of 100 and 190 Hz, respectively) were used to create a two-talker babble. A randomly selected babble segment started 150 ms before and ended 150 ms after each target sentence. The noise level was adjusted to achieve a +10 dB signal-to-noise ratio (SNR). In a pilot study, this SNR yielded medium sentence-recognition scores for the CI and CI+HA conditions (40% and 60% correct, respectively).

TABLE I. Demographic information, fundamental frequency, sentence duration, and intelligibility for each talker in the lexically controlled sentences.

Talker	Talker set	Gender	Age (Yr)	Race/ethnicity	Primary residence	F0 (Hz)				Mean sentence duration (s)	Intelligibility (8 channels) (%)
						Minimum	Maximum	Range	Mean		
T1	-	F	25	Asian	Indiana	63	410	348	253	2.2	67.50
T6	A	F	21	White	Midwest	67	408	341	205	2.5	79.17
T7	A	F	36	White	Illinois	64	412	348	198	2.4	76.11
T3	A	M	27	White	Utah	75	306	231	136	2.6	75.83
T5	A	M	28	White	California	73	258	184	120	2.4	64.72
T2	B	F	23	White	Arizona	66	391	326	210	2.4	76.94
T9	B	F	24	Black/ African-American	New Jersey	66	414	348	192	2.2	77.50
T4	B	M	20	White	Midwest/ East Coast	73	258	184	120	2.4	68.33
T8	B	M	28	Black/ African-American	Indiana	66	391	326	210	2.4	73.61
T10	-	M	27	Black/ African-American	Indiana	66	414	348	192	2.2	50.00

C. Speech processing

1. CI simulation

After pre-emphasis (first-order Butterworth high-pass filter at 1200 Hz), the input speech signal (with or without noise) was band-pass filtered into eight channels (fourth-order Butterworth filters). The overall input acoustic frequency ranged from 100 to 6000 Hz. The corner frequencies of the analysis bands were calculated in accordance with Greenwood's (1990) equation. The band-pass filtered speech was half-wave rectified and low-pass filtered (fourth-order Butterworth filter at 500 Hz) to extract the temporal envelope. The extracted envelope was used to modulate a wide-band noise carrier and filtered using the same pass-bands as the analysis filters. The output of each pass-band was summed to produce the CI simulation. A noise-band vocoder was used instead of a sine-wave vocoder because the spectral side-band cues in sine-wave vocoders are not available to real CI users and may over-estimate their talker-identification performance (Gonzalez and Oliver, 2005). This classical simulation technique (Shannon *et al.*, 1995) mimics the continuous interleaved sampling strategy (CIS; Wilson *et al.*, 1991). Although current CI systems use different processing strategies based on peak picking or current steering, speech performance is generally similar to that with the CIS strategy. For the relatively difficult, lexically controlled VAST sentences, the eight-channel CI simulation yielded moderate sentence-recognition scores similar to the preliminary real CI data from an ongoing study, which also prevented floor or ceiling effects.

2. HA simulation

Speech stimuli (with or without noise) were low-pass filtered using a 10th-order Butterworth filter at 500 Hz (−60 dB/octave) to simulate low-frequency acoustic hearing. This cut-off frequency is higher than human voice F0 and represents the higher end of low-frequency residual hearing typically found in EAS users and bimodal listeners (Dorman *et al.*, 2005; Dorman *et al.*, 2008). Such a HA simulation has been reported in EAS literature (Dorman *et al.*, 2005; Kong and Carlyon, 2007; Brown and Bacon, 2009a). Although it does not account for elevated thresholds, reduced dynamic range, or broadened filters in the low-frequency residual hearing of actual CI users, it provides a theoretical basis for evaluating the usefulness of low-frequency acoustic cues in our training paradigm.

3. F0-controlled sine wave

To examine how much F0 alone contributes to talker identification with simulated bimodal hearing, listeners in the CI+HA group also were tested on talker identification in a CI+F0 condition (Brown and Bacon, 2009a). In this condition, the HA simulation was replaced with an F0-controlled sine wave. The F0 of the target sentence was extracted using an autocorrelation-based algorithm and was used to modulate the frequency of a sine wave, keeping amplitude constant. For the CI+F0 condition, background noise was present in the CI simulation but was not added to the F0-controlled sine wave to maintain robustness of the target F0 cue. In all instances of signal processing, the processed signal was always normalized to have the same long-term root-mean-square (RMS) amplitude as the original unprocessed speech stimuli.

D. Apparatus and procedure

Training and testing were completed in a double-walled sound-attenuating booth and stimuli were presented via headphones (Sennheiser HD 265 linear) at 70 dB SPL in each ear, close to the average conversational level of speech.

Figure 1 shows the training and testing protocol adapted from Nygaard and Pisoni (1998) for subjects in the two experimental groups. Each subject participated in practice and pre-training tests on the first day. On each of the four following days, they participated in training sessions consisting of talker familiarization, talker-identification training, and talker-identification testing, in that order. Training sessions were typically held on four consecutive days (± 2 days). These training days will be referred to as Day 1, Day 2, Day 3, and Day 4, respectively. Subjects completed post-training sentence- and emotion-recognition tests on Day 4 and another sentence-recognition test about 30 days (± 4 days) after Day 4 (1-month post-training). Each training session lasted 45–60 min, making the total training time 4 h, and the total duration of the study (including pre- and post-tests) 5–6 h. The number of training sessions and the duration for each were dictated by the number of sentences in our lexically controlled speech database.

1. Practice

The purpose of the practice tests was to familiarize subjects with the CI simulation and sentence-recognition task. Subjects listened to and orally repeated IEEE sentences

FIG. 1. Schematic of the 4-day talker-identification training protocol for experimental subjects.

processed by the same CI simulation as in the actual study. Subjects were administered two to five lists of sentences until their key-word percent correct in this task reached an asymptote, i.e., performance between two sequential lists did not differ by more than 5%. A Mann–Whitney test showed that asymptotic performance in the practice session did not differ between the experimental CI and CI+HA groups (non-parametric test: $U = 65.50$, $P = 0.73$), suggesting that they were matched in terms of their ability to understand CI simulation speech.

2. Pre-training test

Sentence-recognition performance was assessed using two lists of sentences (in quiet and in noise) presented in random order. Each list consisted of 32 unique sentences (8 talkers $\times$ 4 lexical categories). Subjects were instructed to listen to each sentence and type in what they heard via a keyboard. Emotion-recognition performance was assessed using 100 tokens in quiet, which were presented in random order without repetition. A closed-set, five-alternative forced-choice task was used without feedback.

3. Training

Talker familiarization consisted of listening to 96 tokens (12 sentences $\times$ 4 talkers $\times$ 2 repeats) from the four talkers in the assigned talker set. The same sentences were used for each of the four talkers to avoid a linguistic confound and instead encourage listeners to focus on the talkers' voice. A monitor displayed the static faces of the four talkers in the talker set, each face always paired with an arbitrary and false name that was consistent with gender. Subjects clicked on a talker's face to listen to a sentence from that talker. During talker-identification training, the same sentences were used as in familiarization, but the presentation order was randomized. Subjects listened to the sentences and identified the talkers by clicking on the corresponding face on the monitor and were given auditory as well as orthographic feedback. The sentence was replayed regardless of whether the response was correct or incorrect. The talker-identification test consisted of two lists of 64 tokens (16 novel sentences $\times$ 4 talkers per list), presented in random order and without feedback. For subjects in the CI group, both lists were presented in the CI condition. For subjects in the CI+HA group, one list was presented in the CI+HA condition and the other was presented in the CI+F0 condition; the list order was randomized. Forty-eight sentences used in the pre-training sentence-recognition test were reused during the talker familiarization and training

sessions with a different subset of 12 sentences presented on each of the four training days. This reuse of sentences was necessary due to the limited number of available sentences. However, care was taken to use novel sentences for the talker-identification test at the end of each training day.

4. Post-training test and one-month post-training test

Sentence-recognition tests on Day 4 and 1 month later were carried out in quiet and in noise, each using 32 novel sentences not heard during earlier testing or training. The emotion-recognition test on Day 4 used the same 100 emotional tokens administered before training. Emotion recognition was not re-tested 1 month later due to time constraints.

5. "No training" control group

The greatest benefits from talker-identification training were noted in the CI+HA condition for sentence recognition with Talker Set A (see Sec. III for details). Therefore the control group was only tested for this condition and talker set. Control subjects participated in two sessions that were 4 days apart. The first session consisted of practice, followed by sentence-recognition tests both in quiet and in noise. The second session also consisted of sentence-recognition tests but used novel sentences.

III. RESULTS

The rationalized arcsine transform (Studebaker, 1985) was applied to all talker-identification and sentence-recognition scores prior to statistical analysis to better compare scores at the upper and lower ends of the percent scale. Linear mixed models (SPSS, version 18) were used to analyze the effect of training on talker identification and sentence recognition separately. Bonferroni corrections were used for *post hoc* *t*-tests to follow up significant main effects and interactions. As the emotion-recognition data did not show ceiling or floor effects, they were not subject to the rationalized arcsine transformation. Emotion-recognition scores were analyzed using a two-way analysis of variance (ANOVA).

A. Talker identification

Figure 2 shows the percentage correct talker-identification scores as a function of training day for each of the three processing conditions. Participants in all three processing conditions showed improvement in talker-identification accuracy with training. The greatest improvement from Day 1 to Day 4

FIG. 2. Talker-identification accuracy in the test as a function of training day for each of the three processing conditions (far left panel: CI, middle panel: CI+HA, right-most panel: CI+F0). Mean performance for each talker set is shown separately (set A, filled circles and solid lines; set B, filled triangles and dashed lines). The dashed line through all the three panels represents chance performance. Error bars denote standard error.

TABLE II. Between-gender confusions for subjects in the three processing conditions as a function of training day and talker set. Mean percentage correct responses are in bold.

Processing	Talker set	Gender presented	Gender selected							
			Day 1		Day 2		Day 3		Day 4	
			Female	Male	Female	Male	Female	Male	Female	Male
CI	A	Female	84	16	86	14	88	12	89	11
		Male	11	89	11	89	5	95	7	93
	B	Female	94	6	97	3	96	4	99	1
		Male	7	93	6	94	2	98	4	96
CI+HA	A	Female	100	0	100	0	99	1	100	0
		Male	0	100	1	99	0	100	0	100
	B	Female	100	0	100	0	100	0	100	0
		Male	0	100	0	100	0	100	0	100
CI+F0	A	Female	98	2	100	0	99	1	100	0
		Male	2	98	0	100	1	99	0	100
	B	Female	100	0	100	0	100	0	100	0
		Male	0	100	0	100	1	99	0	100

was noted in the CI+HA condition (17%), followed by the CI+F0 condition (14%) and the CI condition (10%). The main effect of processing was significant [$F(2,37) = 48.21$, $P < 0.001$] with overall talker-identification accuracy being significantly higher for the CI+HA than for the CI condition ($P = 0.01$). Performance in the CI+HA condition also was significantly better than that for the CI+F0 condition ($P < 0.001$). Talker set also affected talker-identification accuracy significantly [$F(1,19) = 8.30$, $P = 0.01$]. Overall talker-identification performance for Talker Set B was significantly better than that for Talker Set A. Finally, training day also had a significant effect on talker-identification accuracy [$F(3,100) = 18$, $P < 0.001$]. Overall performance on Days 3 and 4 was significantly higher than that on Day 1 ($P < 0.001$); Day 4 performance also was significantly better than that on Day 2 ($P < 0.001$).

Linear mixed models and *post hoc* Bonferroni *t*-tests also were used to analyze separately between- and within-

gender confusion matrices generated from talker-identification data. Table II shows the mean between-gender confusion matrix for subjects assigned to each processing condition and talker set as a function of training day. Mean gender recognition in the CI condition was significantly worse [$F(1,40) = 35.40$, $P < 0.001$] than that in the CI+F0 or CI+HA condition. Gender recognition did not differ significantly between the CI+HA and CI+F0 conditions, both of which were at ceiling.

Table III displays mean within-gender confusions for subjects assigned to Talker Set A; similar patterns of confusions were noted for Talker Set B. Participants in the CI or CI+F0 conditions made more within-gender errors than those in the CI+HA condition ($t = -2.60$, $P = 0.01$). However, within-gender errors were similar for subjects in the CI and CI+F0 conditions. Participants in the CI+F0 or CI+HA condition showed similar improvement in within-gender recognition across training days. Their improvements were

TABLE III. Within-gender confusions for subjects trained on talker set A in the three processing conditions as a function of training day. The gender of the talker is indicated within parenthesis (M for male, F for female) and mean percentage correct responses are in bold.

Processing	Talker presented	Talker selected															
		Day 1				Day 2				Day 3				Day 4			
		T3(M)	T5(M)	T6(F)	T7(F)	T3(M)	T5(M)	T6(F)	T7(F)	T3(M)	T5(M)	T6(F)	T7(F)	T3(M)	T5(M)	T6(F)	T7(F)
CI	T3(M)	61	29	1	9	65	27	1	7	66	28	0	6	64	28	1	7
	T5(M)	33	54	1	13	34	52	2	12	34	61	0	5	33	58	1	8
	T6(F)	1	3	64	33	2	2	67	29	2	3	65	31	0	1	72	27
	T7(F)	9	17	20	55	10	15	21	54	9	8	24	58	9	9	25	57
CI+HA	T3(M)	77	23	0	0	83	17	0	0	85	15	0	0	90	10	0	0
	T5(M)	26	74	0	0	22	77	0	1	20	80	0	0	17	83	0	0
	T6(F)	0	0	88	13	0	0	93	7	1	0	96	3	0	0	97	3
	T7(F)	0	0	10	90	0	0	7	93	0	0	3	97	0	0	2	98
CI+F0	T3(M)	57	42	0	1	71	29	0	0	64	36	0	0	79	21	0	0
	T5(M)	26	71	1	2	27	73	0	0	30	69	0	1	18	82	0	0
	T6(F)	0	0	84	16	0	0	83	17	0	0	90	10	0	0	93	7
	T7(F)	2	1	18	79	0	0	16	84	1	0	23	76	0	0	18	82

FIG. 3. Sentence-recognition scores for the CI (left) and CI+HA (right) groups as a function of testing session (pre-, post-, and 1-month post-training) and noise condition (quiet and noise). Performance for familiar/trained talkers is represented by the dark gray bars, whereas performance for unfamiliar/untrained talkers is represented by the light gray bars. Error bars indicate standard error.

significantly higher compared to those of the CI group ($t = 2.87$, $P = 0.004$), who showed little change in within-gender recognition with training.

B. Sentence recognition

Figure 3 shows pre- and post-training sentence-recognition performance for the experimental (CI and CI+HA) groups. Mean sentence-recognition performance in quiet for the CI group was 64% and 82% correct for pre- and post-training, respectively, compared to 83% and 91% correct for the CI+HA group. In the presence of background noise, mean performance for the CI group was 22% and 44% correct pre- and post-training, respectively, compared to 51% and 72% correct for the CI+HA group.

Overall sentence-recognition performance for the CI+HA group was significantly better than that for the CI group [$F(1,20) = 58$, $P < 0.001$]. The addition of background noise significantly reduced sentence-recognition performance [$F(1,208) = 1102$, $P < 0.001$]. Sentence-recognition performance also differed significantly across sessions [$F(2,209) = 134$, $P < 0.001$]; performance measured both immediately and 1 month after training was significantly better than that in the pre-training test. Sentence-recognition performance measured immediately post-training did not differ from that measured a month later.

The two-way interaction between group and noise condition was significant [$F(1,208) = 24$, $P < 0.001$]. Adding noise degraded sentence-recognition performance more for the CI group (40% decreases) than for the CI+HA group (25% decreases). The interaction between noise condition and session also was significant [$F(2,208) = 4.15$, $P = 0.02$]. Noise degraded sentence-recognition performance more in the pre-training test than in either of the post-training tests. Finally, the interaction between talker familiarity and talker set was significant [$F(1,208) = 6.08$, $P = 0.01$]. Subjects trained on Talker Set A demonstrated better sentence recognition (both pre- and post-training) for familiar talkers (Set A) than for unfamiliar talkers (Set B) ($P = 0.03$). However, no such difference was found for subjects trained on Talker Set B ($P = 0.18$). Notably, the interaction between session and talker familiarity was not significant [$F(2,208) = 0.96$, $P = 0.39$], suggesting that

sentence recognition improved similarly for both familiar and unfamiliar talkers from pre- to post-training tests (i.e., a lack of talker familiarity effect).

Linear regression was used to examine the relationship between improvements in talker identification and sentence recognition for each talker. Separate analyses were conducted for the CI and CI+HA groups in quiet and in noise. No significant correlations were noted with the exception of a weak negative correlation between improvements in talker identification and those in sentence recognition in noise for the CI group ($r = -0.29$, $P = 0.05$).

Figure 4 shows sentence-recognition scores for the control and experimental groups, both tested in the CI+HA condition with Talker Set A. A two-way repeated-measures ANOVA showed that sentence-recognition performance for the control group did not vary significantly across sessions [$F(1,13) = 0.01$, $P = 0.94$], but significantly degraded with the addition of noise [$F(1,11) = 44.35$, $P = 0.01$]. Two-way ANOVAs were used to compare the sentence-recognition performance between the control and

FIG. 4. Bar graph comparing sentence-recognition scores for subjects in the control group (solid bars) with those in the corresponding experimental (bars with hash pattern) group. Scores are plotted as a function of testing session and noise condition. Sentence-recognition performance at pre-training (experimental group) or for the first session (control group) is indicated by light shading and that at post-training (experimental group) or for the second session (control group) is denoted by dark bars.

experimental groups in quiet and in noise separately. In quiet, sentence-recognition performance for the experimental group (87% correct) was significantly better [$F(1,30)=5.95$, $P=0.02$] than that for the control group (67% correct), although the effect of session was not statistically significant. In noise, sentence-recognition performance for the experimental group (62% correct) was significantly better [$F(1,30)=4.36$, $P=0.05$] than that for the control group (48% correct). In noise, there was also a significant effect of session, i.e., post-training (or second-session) sentence-recognition performance was significantly better [$F(1,30)=5.46$, $P=0.03$] than pre-training (or first-session) performance. A significant interaction between group and session [$F(1,30)=5.04$, $P=0.03$] also was noted in noise with the experimental group showing more improvement in sentence recognition with training (20% increases) than the control group (5% increases).

C. Emotion recognition

Figure 5 shows pre- and post-training emotion-recognition performance for the experimental groups. Before talker-identification training, the CI group correctly identified 44% of the target emotions, which increased to 49% after training. For the CI+HA group, emotion-recognition performance was 67% correct before training and 77% correct after training. A two-way ANOVA showed that post-training performance was significantly better than pre-training performance [$F(1,44)=8.28$, $P=0.01$]. Performance also was significantly better for the CI+HA group than for the CI group [$F(1,44)=80.19$, $P<0.001$]. However, the interaction between group and session was not statistically significant [$F(1,44)=0.81$, $P=0.37$].

Correlation analyses also were used to examine the relationships among improvements with training in talker identification, sentence recognition, and emotion recognition for individual subjects. Separate analyses were conducted for each group and noise condition. No significant correlations were noted, except that improvement in emotion recognition were positively correlated with those in sentence recognition in noise for the CI group ($r=0.59$, $p=0.04$).

FIG. 5. Mean emotion-recognition scores obtained at pre- and post-training sessions are plotted for the two experimental groups. Pre-training test scores for both groups are represented by filled black bars, whereas scores for the post-training are represented by bars filled with a hash pattern. Error bars indicate standard error.

IV. DISCUSSION

A. Talker identification

The training paradigm used in this study significantly improved the ability of subjects in both experimental groups to identify talkers. However, subjects in the CI+HA group showed significantly better talker-identification performance, suggesting that fine spectral details contained in the low-frequency acoustic signal aided voice identification. Talker-identification accuracy also greatly improved when the CI simulation was combined with an F0-controlled sine wave. However, the CI+F0 performance remained poorer than that in the CI+HA condition. This suggests that in addition to F0, other cues in the 500-Hz low-pass filtered speech such as higher harmonics, the first formant (F1), nasal spectrum cues, and amplitude envelope also may play a role in talker identification. Other researchers have found that F0 alone provided significant benefits, while amplitude envelope and voicing cues also contributed to the sentence-recognition advantage in both simulated and real EAS users (e.g., [Brown and Bacon, 2009a,b](#)). Current widely used implant strategies such as ACE and HiRes ([Vandali et al., 2005](#); [Koch et al., 2004](#)) only convey implicit pitch cues via temporal envelopes, which are insufficient for robust pitch perception. In light of the importance of F0 cues to talker identification and sentence recognition, future implant strategies should provide greater low-frequency spectral resolution and better representation of voice pitch.

Subjects in the CI+HA group showed greater improvement for talker-identification accuracy across training days than subjects in the CI group (see Fig. 2). In the CI+F0 condition, subjects showed intermediate improvement compared to the other two conditions. The trajectory of improvement for the CI and CI+F0 conditions suggests that their performance may continue to improve (although slowly) with a longer training period. However, ceiling effects were noted for the CI+HA condition at the end of Day 4, leaving little room for further improvement. Future investigations may examine the effectiveness of longer-term training in the CI and CI+F0 conditions to further improve training effects.

As mentioned previously, two other studies have examined talker-identification training either in actual CI users ([Barker, 2006](#)) or using CI simulation ([Loebach et al., 2008](#)). [Barker \(2006\)](#) trained CI users using sentence materials and reported a baseline talker-identification performance of 49% correct for 6 talkers and an improvement of roughly 9% after training. Similar improvements were noted in the current simulation study, with fewer talkers ($n=4$) and a training duration that was four times longer. The effectiveness of shorter-term training in the Barker study may be attributed in part to the use of CI subjects with superior speech perception skills. [Loebach et al. \(2008\)](#) trained normal-hearing listeners to identify six talkers using sine-wave CI simulations and showed somewhat greater improvement (12%) for their subjects, although their baseline performance was lower than the current study (42% and 65% correct, respectively). Their use of a sine-wave vocoder may have preserved frequency-modulation information and side-band cues, resulting in

greater improvement in talker identification (Gonzalez and Oliver, 2005).

In the present study, the addition of low-pass filtered speech or F0-controlled sine wave to CI simulation greatly improved between- and within-gender talker identification in contrast to findings of Dorman *et al.* (2008). The different results of the two studies may be in part attributed to different testing materials. In the current study, subjects listened to naturally spoken sentences and therefore had access to sentence intonation, prosody, and speaking rate. In contrast, subjects in the Dorman *et al.* (2008) study listened to word-length materials and would not have had access to sentence-level cues.

B. Sentence recognition

Significant and similar gains in sentence recognition were noted for both experimental groups at the end of talker-identification training in quiet as well as in noise. Talker identification relies on acoustic cues such as voice pitch, harmonic structure, formants, and speaking rate. Some of these cues, such as voice pitch, harmonic structure, and formants are equally relevant to sentence recognition, particularly in degraded conditions such as in a CI simulation and in the presence of competing talkers. We hypothesize that access to these cues during talker-identification training may help subjects better segregate and recognize target speech from a competing-talker background as well. This is consistent with the hypothesis of Kong *et al.* (2005) that the correlation between the salient temporal fine structure cues in low-frequency acoustic hearing and the weak temporal envelope cues in electric hearing may enhance segregation of signal and noise.

It was expected that talker-specific voice learning would lead to improvement in individual subjects' ability to understand the same talker. However, linear regression analyses did not find significant correlations between improvements in these two measures consistent with the findings that improvements in sentence recognition were noted not only for the talkers that subjects were trained to identify but also for novel talkers. Barker (2006) also noted such a generalization of training effects across talkers in CI users. However, Nygaard and Pisoni (1998) reported "talker-specific" improvements in sentence recognition for normal-hearing listeners. In their study, the talker familiarity effect was reported as the difference in post-training sentence recognition between experimental and control groups with a talker set that was familiar only to the experimental group. In contrast, our counterbalanced study design allowed us to compare the relative improvements from pre- to post-training tests for both familiar and novel talkers.

The addition of noise to the speech signal degraded overall sentence recognition to a lesser extent for the CI+HA group than for the CI group. Better sentence recognition in background noise, particularly in the presence of competing talkers, has been reported in simulated bimodal hearing as well as for actual bimodal listeners when compared to performance with CI alone (e.g., Turner *et al.*, 2004). It was further noted that noise affected the CI+HA sentence-recognition perform-

ance to a lesser extent in the post-training than in the pre-training test, suggesting that talker-identification training with low-frequency acoustic cues increased the robustness of speech perception in background noise.

Finally, it may be argued that the improvements in sentence recognition for the experimental groups could be attributed to procedural learning or repeated practice with the testing procedure. To minimize the effects of such learning, practice IEEE lists were administered to subjects before training and testing. Additionally, the similar performance across the two testing sessions in the "no-training" control group suggests that the improvements in the experimental group were not due to test-retest effects. Instead such improvements may be due to the amount of exposure or the specific training with feedback. To examine the role of training with feedback (and that of exposure), three subjects underwent talker-identification training in the CI+HA condition without feedback. With the same amount of exposure as the experimental group, this group only showed an 11% improvement for sentence recognition in noise, similar to the no-training control group. These preliminary results suggest that the sentence-recognition improvements in noise (22%) observed in the experimental group cannot be accounted for simply by exposure but that targeted training with feedback improved perceptual learning.

C. Emotion recognition

Generalization of talker-identification training to emotion recognition, even with a group of novel talkers, suggests that perceptual dimensions that are relevant for talker identification, such as voice pitch, are also shared with emotion recognition. Thus talker-identification training may help listeners better understand implicit emotional information in addition to explicit linguistic information conveyed by the speech signal. The importance of voice pitch and fine-structure cues to emotion recognition has been demonstrated in literature (Luo *et al.*, 2007; Most and Peled, 2007) and is consistent with our findings that the CI+HA group showed an advantage for emotion recognition over the CI group.

D. CI applications

Our simulation results suggest that talker-identification training was effective in simulated electric hearing and more so in the presence of low-frequency acoustic cues. Also its training effects generalized to sentence recognition in quiet and in noise as well as emotion recognition. However, there are inherent differences between real and simulated electric hearing and the large variability in CI users such as the degree of frequency-place mismatch, neural survival, and residual hearing cannot be accounted for by our simulation method. Thus the efficacy of talker-identification training is yet to be seen in actual CI users. Some evidence can be found in Barker (2006), although only unilateral CI users were tested. We have preliminary results from four experienced adult CI users: one EAS (Hybrid), one bimodal, and two unilateral CI users, who completed our talker-identification training protocol. Subjects received variable

benefits in talker identification and sentence recognition. Three of the four subjects showed minor improvements in talker identification (3-8% increases) with training. For sentence recognition in quiet, two of the four subjects showed improvements: 22% increase for the EAS user and 7% increase for one of the unilateral CI subjects. In noise, the EAS user showed no improvement, but the (same) unilateral CI user showed significant improvement (37% increase). The effect of training in CI users is inconclusive based on these preliminary data, but the results underscore the need for follow-up studies to investigate individual variability.

ACKNOWLEDGMENTS

The data presented in this paper are from the dissertation research conducted by the first author in fulfillment of requirements for the doctoral degree at Purdue University. This research was supported, in part, by the National Institutes of Health (NIDCD Grant Nos. T32 DC000030, R01 DC008875-03, and R03 DC008192-04). We thank Ching-Chih Wu for technical assistance with programming and Larry Humes for feedback on an earlier version of this manuscript.

- Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **88**, 680–697.
- Barker, B. A., "An examination of the effect of talker familiarity on the sentence recognition skills of cochlear implant users," Ph.D. dissertation, The University of Iowa, Iowa City, Iowa, 2006.
- Bell, T. S., and Wilson, R.H. (2001). "Sentence recognition materials based on frequency of word use and lexical confusability," *J. Am. Acad. Audiol.* **12**, 514–522.
- Brox, J. P. L., and Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phonetics* **10**, 23–36.
- Brown, C. A., and Bacon, S. P. (2009a). "Low-frequency speech cues and simulated electric-acoustic hearing," *J. Acoust. Soc. Am.* **125**, 1658–1665.
- Brown, C. A., and Bacon, S.P. (2009b). "Achieving electric-acoustic benefit with a modulated tone," *Ear Hear.* **30**, 489–493.
- Buchner, A., Schu'sshler, M., Battmer, R. D., Stover, T., Lesinski-Schiedat, A., and Lenarz, T. (2009). "The impact of low-frequency hearing," *Audiol. Neurotol.* **14**, S1, 8–13.
- Dorman, M. F., Gifford, R., Spahr, A. J., and McKarns, S. (2008). "The benefits of combining electric and acoustic stimulation for the recognition of speech, voice, and melodies," *Audiol. Neurotol.* **13**, 105–112.
- Dorman, M. F., Spahr, A. J., Loizou, P. C., Dana, C. J., and Schmidt, J. S. (2005). "Acoustic simulations of combined electric and acoustic hearing (EAS)," *Ear Hear.* **26**, 1–10.
- Fu, Q.-J., and Galvin, J. J. (2007). "Perceptual learning and auditory training in cochlear implant recipients," *Trends Amplif.* **11**, 193–205.
- Fu, Q.-J., Galvin, J. J., Wang, X., and Nogaki, G. (2005). "Moderate auditory training can improve speech performance of adult cochlear implant patients," *ARLO* **6**, 106–111.
- Gonzalez, J., and Oliver, J. C. (2005). "Gender and speaker identification as a function of the number of channels in spectrally degraded speech," *J. Acoust. Soc. Am.* **118**, 461–470.
- Green, T., Faulkner, A., and Rosen, S. (2002). "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," *J. Acoust. Soc. Am.* **112**, 2156–2164.
- Greenwood, D. D. (1990). "A cochlear frequency-position function for several species- 29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Institute for Electrical and Electronic Engineers. (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio. Electro Acoust.* **17**, 225–246.
- Koch, D. B., Osberger, M. J., Segel, P., and Kessler, D. K. (2004). "HiResolution and conventional sound processing in the HiResolution Bionic Ear: Using appropriate outcome measures to assess speech-recognition ability," *Audiol. Neurotol.* **9**, 214–223.
- Kong, Y.-Y., and Carlyon, R. P. (2007). "Improved speech recognition in noise in simulated binaurally combined acoustic and electric simulation," *J. Acoust. Soc. Am.* **121**, 3717–3727.
- Kong, Y.-Y., Stickney, G. S., and Zeng, F.-G. (2005). "Speech and melody recognition in binaural combined acoustic and electric hearing," *J. Acoust. Soc. Am.* **117**, 1351–1361.
- Kreiman, J., and Sidtis, D. V. L. "Physical characteristics and the voice: can we hear what a speaker looks like? in *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell Publishing, West Sussex, UK, 2011), Chap. 4, pp. 110–155.
- Loebach, J. L., Bent, T., and Pisoni, D. B. (2008). "Multiple routes to the perceptual learning of speech," *J. Acoust. Soc. Am.* **124**, 552–561.
- Luo, X., Fu, Q.-J., and Galvin, J. (2007). "Vocal emotion recognition by normal-hearing listeners and cochlear-implant users," *Trends Amplif.* **11**, 301–315.
- Most, T., and Peled, M. (2007). "Perception of suprasegmental features of speech by children with cochlear implants and children with hearing aids," *J. Deaf Stud. Deaf Educ.* **12**, 350–361.
- Murray, I. R., and Arnott, J. L. (1993). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Am.*, **93**, 1097–1108.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophy.* **60**, 355–376.
- Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). "Speech perception as a talker-contingent process," *Psychol. Sci.* **5**, 42–46.
- Oxenham, A. J. (2008). "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," *Trends Amplif.* **12**, 316–331.
- Remez, R. E., Fellowes, J. M., and Rubin, P. E. (1997). "Talker identification based on phonetic information," *J. Exp. Psychol.: Hum. Percept. Perform.* **23**, 651–666.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assman, P. F. (2004). "Cochlear implant speech recognition with speech maskers," *J. Acoust. Soc. Am.* **116**, 1081–1091.
- Straatman, L. V., Rietveld, A. C. M., Beijen, J., Mylanus, E. A. M., and Mens, L. H. M. (2010). "Advantage of bimodal fitting in prosody perception for children using a cochlear implant and a hearing aid," *J. Acoust. Soc. Am.* **128**, 1884–1895.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (2004). "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," *J. Acoust. Soc. Am.* **115**, 1729–1735.
- Vandali, A. E., Sucher, C., Tsang, D. J., McKay, C. M., Chew, J. W., and McDermott, H. J. (2005). "Pitch ranking ability of cochlear implant recipients: A comparison of sound-processing strategies," *J. Acoust. Soc. Am.* **117**, 3126–3138.
- von Ilberg, C., Kiefer, J., Tillein, J., Pfenningdorff, T., Hartmann, R., Sturzebecher, E., and Klinke, R. (1999). "Electricacoustic stimulation of the auditory system: New technology for severe hearing loss," *ORL J. Otorhinolaryngol. Relat. Spec.* **61**, 334–340.
- Vongphoe, M., and Zeng, F.-G. (2005). "Speaker recognition with temporal cues in acoustic and electric hearing," *J. Acoust. Soc. Am.* **118**, 1055–1061.
- Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (1991). "Better speech recognition with cochlear implants," *Nature* **352**, 236–238.