Voice gender and the segregation of competing talkers: Perceptual learning in cochlear implant simulations

Jessica R. Sullivan,^{1,a)} Peter F. Assmann,² Shaikat Hossain,² and Erin C. Schafer³

¹*Department of Communication Sciences & Professional Counseling, University of West Georgia, Carrollton, Georgia 30118, USA*

²*School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, Texas 75083, USA*

³*College of Health and Public Service, University of North Texas, Denton, Texas 76203, USA*

(Received 13 August 2014; revised 26 December 2016; accepted 27 January 2017; published online 9 March 2017)

Two experiments explored the role of differences in voice gender in the recognition of speech masked by a competing talker in cochlear implant simulations. Experiment 1 confirmed that listeners with normal hearing receive little benefit from differences in voice gender between a target and masker sentence in four- and eight-channel simulations, consistent with previous findings that cochlear implants deliver an impoverished representation of the cues for voice gender. However, gender differences led to small but significant improvements in word recognition with 16 and 32 channels. Experiment 2 assessed the benefits of perceptual training on the use of voice gender cues in an eight-channel simulation. Listeners were assigned to one of four groups: (1) word recognition training with target and masker differing in gender; (2) word recognition training with same-gender target and masker; (3) gender recognition training; or (4) control with no training. Significant improvements in word recognition were observed from pre- to post-test sessions for all three training groups compared to the control group. These improvements were maintained at the late session (one week following the last training session) for all three groups. There was an overall improvement in masked word recognition performance provided by gender mismatch following training, but the amount of benefit did not differ as a function of the type of training. The training effects observed here are consistent with a form of rapid perceptual learning that contributes to the segregation of competing voices but does not specifically enhance the benefits provided by voice gender cues. © 2017 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4976002>]

[DB]

Pages: 1643–1652

I. INTRODUCTION

Over the past 20 years, cochlear implants (CI) have emerged as a viable option to provide auditory access to individuals who receive little to no benefit from acoustic hearing aids. In general, CI users' speech perception performance and the quality of CI technology have improved (Gifford and Revit, 2010). Despite the remarkable progress CI users have shown in understanding speech in quiet, there remains a great deal of variability in their performance, especially in adverse listening situations. The source of such variability can be attributed to a variety of factors related to the implant device (Bierer and Faulkner, 2010) or the listener (Geers *et al.*, 2002; Skinner, 2003; Tobey *et al.*, 2003). Duration of auditory deprivation, etiology of deafness and age of implantation are factors that can affect an individual's speech perception abilities. The difficulties faced by CI users in challenging conditions, such as in the presence of a competing talker, appear to be closely linked to degradation of the spectral and temporal properties of speech, including the cues that convey information about voice gender and pitch (Chatterjee and Peng, 2008; Fuller *et al.*, 2014; Fu *et al.*, 2004a; Stickney *et al.*, 2004).

In listeners with normal hearing, differences between males and females in fundamental frequency (F_0) and formant frequencies, FFs (or relational properties of the spectrum envelope) provide important cues for voice gender perception (Smith and Patterson, 2005; Hillenbrand and Clark, 2009; Skuk and Schweinberger, 2014). Men have lower FFs and lower F_0 s than women, on average, due in large part to differences in the length of the vocal tract and vocal folds, respectively (Fitch and Giedd, 1999; Smith and Patterson, 2005). The ability to recognize and discriminate voices by gender is an important skill used daily in a variety of listening environments, particularly when there are multiple talkers. For listeners with normal hearing, a target talker is more intelligible when a competing talker differs in F_0 (Brox and Nooteboom, 1982) and when competing voices differ in formant frequencies (Darwin *et al.*, 2003; Başkent and Gaudrain, 2016), consistent with the finding that differences in voice gender make it easier to segregate pairs of competing voices (Brungart, 2001). Competing-talker situations are difficult for listeners with normal hearing (NH) but extremely challenging for those with hearing loss, including CI users. While normal hearing listeners are able to benefit from differences in voice gender, this advantage is greatly reduced in CI users, presumably linked to their diminished spectral resolution and/or poorer representation of voice pitch. Stickney *et al.* (2004) measured speech recognition

^{a)}Electronic mail: sulli10@uw.edu

performance in adults with normal hearing (including both unprocessed speech and CI simulation conditions) and CI listeners using maskers consisting of either steady-state speech shaped noise (SSN) or competing sentences spoken by the same male, a different male, or a female talker. While normal hearing listeners in their study performed better with single-talker maskers than with SSN, no difference was observed in the CI or CI simulation groups. Although there was a significant improvement when the target and masker were spoken by different (male) talkers compared to the same male talker, there was no further improvement when the target and masker talkers differed in gender.

Fu *et al.* (2004a) explored the importance of temporal and spectral cues in voice gender identification and vowel recognition by CI users and normal hearing listeners using CI simulations. They found that both voice gender and vowel identification improved as the number of spectral channels was increased for normal-hearing listeners, with similar results for CI users receiving 4–8 spectral channels. Temporal cues were particularly important in identifying voice gender when the number of spectral channels (4–8) was limited. Li and Fu (2011) reported that voice gender discrimination performance was significantly correlated with speech recognition accuracy when temporal envelope cues were present. Gaudrain and Başkent (2015) have shown that spectral envelope cues associated with differences in vocal tract length between male and female voices depend strongly on the number of channels in vocoder simulations. In comparison, F_0 cues are more resilient to reduced spectral resolution, despite the fact that F_0 is poorly transmitted in present-day CI devices (Zeng *et al.*, 2014). Fuller *et al.* (2014) compared the role of F_0 and spectral envelope cues in gender recognition for CI users and listeners with normal hearing, who heard either unprocessed speech or eight-channel vocoder simulations. For the CI users, the cue weight associated with F_0 was much larger than that associated with spectral envelope. In comparison, unprocessed speech resulted in high cue weights for both properties. For the vocoder simulations, listeners relied on both properties but with reduced weights. The authors suggested that the low weights assigned to spectral cues might be related to channel interaction which is not captured effectively in the simulations. On the other hand, they observed highly non-monotonic categorization functions for CI users and simulations, indicating that gender processing is impaired in both groups.

Given that voice gender recognition can be a relatively difficult task for CI users, with considerable variability across listeners (Fu *et al.*, 2004a; Massida *et al.*, 2013), it may be possible to improve speech recognition performance through targeted exposure to gender differences between voices. Aural rehabilitation, in particular, auditory training, is beneficial in improving the speech recognition abilities of CI users and listeners with normal hearing attending to CI simulations (Fu *et al.*, 2004b; Fu and Galvin, 2007; Fu *et al.*, 2005; Nogaki *et al.*, 2007). According to perceptual learning studies, the training task and materials, as well as the duration of the task, determines the degree of immediate and long-term benefit. Several researchers have shown that short-term auditory training can be effective in developing perceptual changes. In particular, auditory training paradigms designed

to adapt to the individual's performance were successful in demonstrating desired training effects (Alain and Tremblay, 2007; Fu and Galvin, 2007; Sullivan *et al.*, 2013; Tremblay, 2007). The present study extends these techniques to determine whether listeners attending to CI simulations can be trained to focus on voice gender cues in the challenging situation where competing talkers are present.

Several studies have demonstrated large improvements in perceptual skills following auditory training via CI simulation. The maintenance and generalization of these skills following training depend on the task and type of stimuli, ranging from environment sounds (Loebach and Pisoni, 2008) to sentences (Davis *et al.*, 2005). Active engagement in training exercises can produce an accelerated improvement in perceptual abilities (Fu and Galvin, 2007). Loebach *et al.* (2008) assessed the efficacy of auditory training on indexical properties of speech (gender and talker recognition) and linguistic properties (word transcription) via CI simulation in adult listeners with normal hearing and found comparable levels of improvement. Voice gender training produced somewhat smaller benefits, possibly because gender recognition in quiet is a relatively easy task.

Krull *et al.* (2012) evaluated the effects of auditory training on a talker identification task in quiet and in babble noise for young adults with normal hearing. Listeners attended to either unilateral or bimodal simulations (CI alone or CI combined with a hearing aid in the non-implanted ear). Talker identification improved with both types, with overall better performance for simulated bimodal hearing. Talker identification training generalized to sentence recognition in quiet and in noise, and this improvement did not depend on whether the sentences were spoken by the talkers heard during training. The findings suggest that training can lead to improvements in speech recognition in noise, can generalize to untrained talkers and materials, and may yield improvements that are maintained over time. However, the study did not explicitly examine the perception of voice gender.

One purpose of the experiments described in this paper was to examine how listeners use voice gender cues in a competing talker situation through CI simulations. While several studies examined how voice gender cues are extracted by CI users and listeners attending to CI simulations (Fu *et al.*, 2004a; Fu *et al.*, 2005; Li and Fu, 2011; Schvartz and Chatterjee, 2012; Visram *et al.*, 2012; Fuller *et al.*, 2014; Gaudrain and Başkent, 2015), many questions remain about how they are used in situations where competing talker are present (Stickney *et al.*, 2004; Qin and Oxenham, 2005; Visram *et al.*, 2012). In the present study, experiment 1 examined the effects of voice gender on speech intelligibility in two-voice combinations where target and masker sentences were either matched in gender (MM, FF) or mismatched (MF, FM). The purpose was to determine conditions in which mismatched target and masker sentences provide a benefit for sentence intelligibility in the presence of a competing talker. Listeners showed a benefit from the mismatched gender of target and masker in 8, 16, and 32 channel conditions, motivating a second experiment to determine if perceptual training on voice gender identification with competing talkers could

lead to an improvement in word recognition accuracy for stimuli similar to those used in experiment 1.

Experiment 2 compared three training conditions: (1) word recognition with target and masker mismatched in gender; (2) word recognition with target and masker matched in gender; (3) gender recognition for sentence pairs mismatched in gender. If perceptual training produces heightened sensitivity to voice gender cues, which in turn facilitates the segregation of competing voices, then each of the three training conditions might lead to improved intelligibility in competing talker conditions. Alternatively, training might be most effective when the training task is closely matched to the evaluation task. This leads to the prediction that mismatched-gender training will be more effective than matched-gender training when tested with a mismatched-gender word recognition task, and that training on word recognition (of either type) will provide greater benefits for word recognition than for gender recognition.

II. SIGNAL PROCESSING

A. Voice processing

Male and female voices differ in several respects, including F_0 , spectrum envelope properties associated with different formant frequency ranges, and additional properties associated with the voicing source (Hillenbrand and Clark, 2009). Studies of voice gender perception by CI users and in CI simulations have used either systematic parametric manipulation of F_0 and spectrum envelope properties in vocoder-processed sentences (e.g., Fuller *et al.*, 2014) or natural voices where cues for voice gender vary considerably across individuals (e.g., Visram *et al.*, 2012). To control for individual differences and to facilitate comparisons across studies, we adopted a simulation approach. Acoustic measurements (F_0 and formant frequencies) were obtained for each of the sentences in the Hearing In Noise Test (HINT; Nilsson *et al.*, 1994), a widely used and carefully controlled sample of 260 sentences produced by an adult male talker. The geometric mean F_0 and formant frequencies (across the three formants) were estimated at 2-ms intervals from all voiced frames. Next, we used a high-quality vocoder (STRAIGHT; Kawahara *et al.*, 1999) to scale the F_0 and spectrum envelope of the HINT sentences to the geometric mean of F_0 and formant frequencies for adult male and female voices in our vowel database (Assmann and Katz, 2000). This approach allowed us to control variables other than F_0 and formant frequencies that may contribute to voice gender. The re-synthesized sentences scaled to the male average were virtually indistinguishable from the original recordings, and those scaled to the female average, were heard as a natural-sounding woman's voice.

B. CI simulations

CI simulations were similar to those described by Fu *et al.* (2004a). First, the signal was processed with a high-pass pre-emphasis filter with a 1200-Hz cutoff and a slope of -6 dB/octave. Next, the frequency region between 200 and 7000 Hz was subdivided using 4, 8, 16 or 32 bandpass filters with center frequencies equally spaced on the Greenwood cochlear

distance function (Greenwood, 1990). Fourth-order Butterworth filters were used, with the corner frequencies listed in Fu *et al.* (2004a, Table II). Envelope extraction was carried out using half-wave rectification and LP filtering (160 Hz cutoff).

The excitation signal was either band-pass filtered white noise (noise vocoders) or a sinusoid (tone vocoders) with the sinusoid positioned at the center frequency of each filter. Several studies have found that sine-wave and noise carriers in CI simulations are comparable in terms of overall speech intelligibility (Dorman *et al.*, 1998; Fu *et al.*, 1998; Loizou *et al.*, 2000; Friesen *et al.*, 2001). On the other hand, some studies have shown that gender recognition is slightly better with sine-wave vocoders (Gonzalez and Oliver, 2005), making it important to include this comparison in experiment 1.

C. Target-masker construction

Each stimulus consisted of a target sentence centered at the temporal midpoint of a speech masker at a target-to-masker ratio of 0 dB. The masker was composed of two concatenated sentences with no intervening silence. The target always started after the onset of the first masker sentence and terminated before the offset of the second masker. The mean time delay between the onset of the masker and the onset of the target was 0.83 s; the minimum delay was 0.56 s. The mean overall duration of the target plus masker signal was 1.58 s. The target plus masker mixture was subsequently processed through a CI simulation. All signal processing was carried out in MATLAB using custom scripts.

III. EXPERIMENT 1: VOICE GENDER DIFFERENCES AND COMPETING TALKERS IN CI SIMULATIONS

The purpose of experiment 1 was to examine sentence intelligibility in the presence of a competing talker as a function of differences in voice gender between target and masker for CI simulations differing in source type (sinusoidal versus noise) and in the number of channels (4, 8, 16, or 32).

A. Method

1. Participants

Sixteen young adults, 4 males and 12 females, participated in this experiment. Participants were native speakers of American English, with no known history of hearing impairment. Hearing was screened at octave frequencies (500, 1000, 2000, and 4000 Hz) ≤ 25 dB hearing level. All were students at the University of Texas at Dallas and received research credits for their participation.

2. Stimuli

Manipulation of the following factors: number of channels (4, 8, 16, 32), voice gender (matched vs mismatched), and source (sinusoidal vs noise) occurred over 128 trials. The 280 HINT sentences were processed as described in Sec. II A to produce male and female versions. A subset of 128 sentences served as targets; the remainder were used as maskers. Each stimulus consisted of a single target sentence

embedded within two consecutive masker sentences as described in Sec. II C. Targets and maskers were paired using random sampling (with the constraint that each sentence could appear only once as a target). In the matched-gender set, the target and maskers were both male (MM) or both female (FF). In the mismatched-gender set, the target differed from the masker sentences in gender (MF or FM). These sentence pairs were subsequently processed to produce eight CI simulations (tone or noise source with 4, 8, 16, or 32 channels).

3. Procedure

All testing was conducted in a sound-treated booth. Stimuli were presented monaurally using Tucker-Davis System 3 hardware through the left channel of Sennheiser 580 circumaural headphones at an average level of 65 dB(A). Prior to testing, participants were given a practice run to familiarize them with the procedure and stimuli. Participants were instructed to listen for the middle (target) sentence and to type the words they heard into a response panel on the computer screen. All sessions were self-paced, participants were allowed breaks and most sessions were completed in less than 1 h.

The stimuli were presented in blocks of eight with the number of channels decreasing from 32 (easiest) to 4 (hardest) across blocks. Matched- and mismatched-gender sentence pairs were randomly interspersed within each block. Half of the listeners completed the noise source condition followed by the sine condition (64 trials each); the other half reversed this order. No repetition of stimuli or feedback was provided during testing.

B. Results and discussion

Speech recognition performance was summarized in terms of the proportion of keywords in the target sentences that were correctly identified. Figure 1 shows word recognition accuracy as a function of number of channels, source type and gender mismatch. Each condition was represented by 16 sentences, with 4 to 6 keywords per sentence. Percent correct scores were transformed to RAU scores (Studebaker, 1985) prior to the statistical analysis. Across all conditions, word recognition accuracy was similar for male and female targets (43.9 and 40.1 RAU, respectively) and the data were combined across male and female targets. A repeated-measures analysis of variance (ANOVA) performed on the RAU scores indicated a significant main effect of the number of channels [$F(3,45) = 234.95$; $p < 0.01$, $\eta^2 = 0.77$] with progressively increasing accuracy as the number of channels increased from 4 to 32. There was also a significant main effect of target-masker mismatch in gender [$F(1,25) = 65.87$; $p < 0.01$, $\eta^2 = 0.19$] with higher scores when target and masker were mismatched in gender, e.g., MF or FM compared to MM or FF combinations, and an interaction effect between the number of channels with gender mismatch [$F(3,45) = 4.59$; $p < 0.01$, $\eta^2 = 0.08$]. Simple effects tests (Bonferroni correction, $\alpha = 0.05$) revealed significantly higher performance with a difference in voice gender for CI simulations with eight or more channels. There was no significant

FIG. 1. Boxplots (medians and interquartile ranges) of word recognition scores (RAU) in experiment 1 for noise vocoders (upper panel) and tone vocoders (lower panel) with matched gender target-masker combinations (dark gray bars) and mismatched combinations (light gray bars) as a function of the number of channels. The “whiskers” represent 1.5 times the inter-quartile range; dots indicate outliers.

effect of source type, nor was there a significant interaction of number of channels by source type.

Overall, performance was better when the target and masker were mismatched (male target combined with female maskers, or vice versa). Sentence recognition with 8, 16, or 32 channels was higher when the masker and target differed in voice gender. Stickney *et al.* (2004) found no differences between gender-matched and gender-mismatched conditions. The discrepancy between studies may be related to differences in the properties of the masker and target sentences. Stickney *et al.* aligned the target and masker sentences at their onsets. To avoid confusion of the target and masker, they used only one masking sentence, which was paired with each of the target sentences. Predictability of the masker might have limited its effectiveness. In addition, the IEEE sentences used in the Stickney *et al.* study are more difficult than the HINT sentences used in the present study and lower in predictability. In comparison, the asynchronous presentation of the target and masker in the current study provides a powerful cue for the onset of the target, directing listeners’ attention to the presence of the second voice.

IV. EXPERIMENT 2: PERCEPTUAL LEARNING AND THE CONTRIBUTION OF VOICE GENDER TO INTELLIGIBILITY WITH COMPETING TALKERS

Experiment 2 assessed the effects of perceptual training with exposure to voice gender cues on word recognition and gender recognition in the presence of a competing talker. Vocoder simulations were constructed using an eight-channel tone vocoder. The vocoder and channel selection were based, in part, on the results of experiment 1, targeting

conditions for which voice gender differences provide an opportunity for improvement, and previous studies comparing simulations with CI users. Gonzalez and Oliver (2005) found that tone vocoders produced more accurate gender recognition than noise vocoders. Fu *et al.* (2004a) found similar performance (but relatively low accuracy) on gender recognition for cochlear implant users and eight-channel vocoder simulations. In experiment 2, both training and test stimuli were similar to those described in experiment 1, with a single target embedded in two masking sentences.

A. Methods

1. General Training

Forty participants were randomly assigned to one of the four groups: word recognition training (matched gender); word recognition training (mismatched gender); masked-gender recognition training (mismatched gender), and control (no training). Participants were asked to complete two training sessions on two separate days within a one-week period. All training and testing was conducted in the same sound booth with similar procedures as described in experiment 1. The training tasks were adapted from custom MATLAB programs and stimuli used by Sullivan *et al.* (2013). The training paradigm was designed to increase difficulty by adapting the TMR and increasing the number of keywords to report as each block progressed. All training tasks were presented in 10 blocks of 18 trials per session (360 total trials across 2 days). All tasks used a computer-controlled training procedure designed to adapt the target-masker ratio (TMR) based on the participant's performance. All training began at 0 dB TMR for each word recognition training group. At the end of each block, stimulus presentation switched to an easier TMR of +6 dB (if the block-averaged word recognition score fell below the criterion level of 70% words correct) or to a more difficult TMR of -3 dB (for scores equal to or higher than 70%). Feedback was provided after each trial in all training conditions. If the trial was correct, a smiley face appeared on the screen, and if the trial was incorrect, a yellow bar was displayed. All scoring was automated and recorded for each session. The instructions were similar for each of the three training groups. Participants were instructed to listen to the middle (target) sentence and then identify the keyword(s) heard using the computer mouse to select a response button from an answer grid that displayed a set of icons representing the available response options along with superimposed printed keywords.

Approximately 1500 sentences were extracted from the Children Auditory Training (CAT) corpus of 3000 sentences used in Sullivan *et al.* (2013). The masker and target sentences were chosen to be thematically different and did not have overlapping target words. Simple declarative sentences were used with the final three words serving as keywords (e.g., *He ate three green grapes*). Two adult native English speakers produced the training sentences, one male and one female. The matched gender condition included both male and female combinations (i.e., male target and male masker as well as female target and female masker) while the mismatched gender condition included both combinations of

male and female targets and maskers. The male talker had a mean F_0 of 82 Hz; the female talker had a mean F_0 of 164 Hz. The female talker's formant frequencies (measured across all voiced frames) were on average 33% higher than the male talker's formant frequencies. In comparison, the male talker in the pre, post and late test sessions had a mean F_0 of 106 Hz, while the female talker had a mean F_0 of 215 Hz. The female talker used in the test sessions had formant frequencies that were on average 13% higher than those of the male talker.

2. Word recognition training

For word recognition training, response buttons were arranged on the answer grid in three columns, each with six rows representing the available choices. The word recognition (mismatched and matched gender) training programs were designed to increase in memory load after every six trials, starting with recalling keywords based on the position in the sentence: (1) last word only: 6 possible choices, (2) middle and last: 12 possible choices, and (3) beginning, middle and last: 18 possible choices. On the computer screen, only the targeted positions of the keyword(s) were visible to participants, and the non-targeted positions were hidden. For example, if the target sentence was "he saw three blue cows" and the target was the last position only, the participant would see only the last column with animal names, including the word "cows." Participants in word recognition training were either given matched or mismatched stimuli according to their assigned condition, with male and female talkers randomly alternating as targets 50% of the time.

3. Gender recognition training

For gender recognition training, listeners were instructed to identify the gender of the target sentence and select one of two response buttons labeled male or female. They received feedback after each trial, similar to the word recognition training. Target-masker combinations included both matched- and mismatched-gender pairs, with male and female targets evenly distributed within each block. Unlike the word-recognition training groups, there was no manipulation of keyword position to increase the memory load, but the procedure for TMR adaptation was the same. Similar to the word recognition training, participants were not given a response time limit; selection of a response initiated the onset of the next trial.

4. Evaluation of training

Test sessions were conducted before the onset of training (pre-test), following the completion of the two training sessions (post-test), and 7 days after the completion of training (late test). Test sessions included a gender identification task as well as speech intelligibility in which participants were asked to type the words heard from the target sentence as described in experiment 1. Each of the three test sessions (pre, post, and late) included 40 sentence pairs for word recognition and 40 for gender recognition, equally divided between matched and mismatched stimuli.

B. Results and discussion

1. Word recognition

Word recognition performance was summarized in terms of the proportion of keywords in the HINT sentences that were correctly identified. Means and standard deviations are shown in Table I. A one-way ANOVA on the pre-test scores did not reveal significant differences among the four groups prior to training, [$F(3,36) = 1.32$; $p = 0.28$]. The effects of training were assessed by computing difference scores (post-test scores minus pre-test scores). Scores were transformed to RAU and an ANOVA was performed on the RAU-transformed difference scores with two factors: (1) training group, a between-subjects factor with four levels: control (no training sessions), matched (word recognition training with same gender target and masker), mismatched (word recognition training with target and masker differing in gender), and gender (gender recognition training with target and masker differing in gender); and (2) gender mismatch, a within-subjects factor with two levels: matched (HINT sentences with target and masker matched in gender) and mismatched (HINT sentences with target and masker differing in gender). Means and standard errors are shown in the upper panel of Fig. 2.

The main effect of training group approached significance [$F(3,36) = 2.81$, $p = 0.053$, $\eta^2 = 0.16$]. Planned comparisons (Bonferroni correction, $\alpha = 0.05$) revealed that training on the word recognition task led to significantly higher difference scores (i.e., improved performance from pre- to post-test) compared to those in the control group who received no training. Listeners who received training on the gender recognition task also showed greater improvements in word recognition post-training compared to those in the control group who received no training, and in fact, training on the word recognition tasks did not provide greater benefits for word recognition than training on the gender recognition task. For all training groups, target plus masker combinations that differed in voice gender (mismatched test stimuli) led to greater benefits of training than same-gender (matched) combinations, leading to a significant main effect of gender mismatch [$F(1,36) = 31.70$, $p < 0.01$, $\eta^2 = 0.13$]. The interaction of training group and gender mismatch was not significant [$F(3,36) = 2.36$; $p = 0.09$, $\eta^2 = 0.03$]. Contrary to predictions, listeners trained on target plus masker combinations that differed in voice gender (mismatched training group) did not

FIG. 2. Boxplots (medians and interquartile ranges) of word recognition difference scores (RAU) between PRE and POST tests (upper panel) and between LATE and PRE tests (lower panel) with matched gender target-masker combinations (dark gray bars) and mismatched combinations (light gray bars) for each training group. The “whiskers” represent 1.5 times the inter-quartile range; dots indicate outliers.

obtain larger benefits from training on mismatched test stimuli than those trained on same-gender combinations (matched training).

To determine whether the training effects were maintained from pre to late conditions, a further analysis of variance was performed on the RAU-transformed difference scores (late-minus pre-test scores) as shown in the lower panel of Fig. 2. As with the pre-post comparisons, improvements in word recognition scores were greater for groups who received training compared to the control group with no training. There was a significant effect of stimulus gender mismatch [$F(1,36) = 9.77$, $p < 0.01$, $\eta^2 = 0.08$] with greater improvement for mismatched stimuli. The interaction of training group and stimulus gender mismatch was not significant.

2. Gender recognition

Gender recognition accuracy is summarized in Table II. A one-way ANOVA on the RAU-transformed pre-test scores did not reveal significant differences among the four groups prior to training [$F(3,36) = 1.94$; $p = 0.14$]. The effects of

TABLE I. Means and standard deviations of word identification accuracy (percent) for experiment 2.

		Control		Gender		Matched		Mismatched	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
PRE	Matched	17.55	8.92	12.58	10.67	15.73	8.38	12.01	8.642
	Mismatched	24.95	10.24	20.32	10.85	15.58	11.29	16.03	10.42
POST	Matched	21.91	8.85	30.16	14.23	27.38	11.84	28.53	10.047
	Mismatched	39.60	9.75	44.34	14.12	43.66	17.27	43.59	10.693
LATE	Matched	29.65	13.82	30.38	14.11	28.89	11.79	31.09	11.999
	Mismatched	42.56	12.79	43.21	14.29	44.08	13.01	44.51	12.033

TABLE II. Means and standard deviations of gender identification accuracy (percent) for experiment 2.

		Control		Gender		Matched		Mismatched	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
PRE		59.5	9.704	50.75	5.898	53.25	10.74	52.5	6.562
POST		60.5	13.218	70.25	12.497	67.5	12.64	78	12.791
LATE		57	10.193	71.75	10.277	70	11.12	75.25	11.634

training were assessed by computing difference scores (post-test scores minus pre-test scores) for each listener. An analysis of variance was performed on the RAU-transformed difference scores with two factors: (1) training group, a between-subjects factor with four levels: control (no training sessions), matched (word recognition training with same gender target and masker), mismatched (word recognition training with target and masker differing in gender), and gender (gender recognition training with target and masker differing in gender); and (2) gender mismatch, a within-subjects factor with two levels: matched (HINT sentences with target and masker matched in gender) and mismatched (HINT sentences with target and masker differing in gender). The results are shown in the upper panel of Fig. 3.

There was a significant main effect of training group [$F(3,36) = 11.81$, $p < 0.01$, $\eta^2 = 0.29$]. Planned comparisons revealed that listeners in each of the training conditions showed greater improvement in gender recognition from pre- to post-test than the control group without training in all three training groups. Gender training led to greater improvement in gender recognition accuracy compared to word recognition training, but the difference did not reach significance. There was a significant effect of gender mismatch [$F(1,36) = 11.78$; $p < 0.01$, $\eta^2 = 0.16$], with higher scores in the mismatched condition, but the interaction of training group and gender mismatch was not significant.

To determine whether the training effects are maintained from pre to late conditions, a further analysis of variance was performed on the RAU-transformed difference scores (late minus pre-test scores) as shown in the bottom panel of Fig. 3. There was a significant main effect of

training group [$F(3,36) = 13.99$, $p < 0.01$, $\eta^2 = 0.36$]. Planned comparisons revealed that listeners in each of the training conditions had higher gender recognition difference scores compared to the control group without training for all three training groups. Gender training did not lead to higher gender recognition difference scores compared to word recognition training. There was a significant main effect of gender mismatch [$F(1,36) = 15.20$, $p < 0.01$, $\eta^2 = 0.19$] and a significant interaction of training group and gender mismatch [$F(3,36) = 4.00$, $p < 0.01$, $\eta^2 = 0.15$]. The gender training group showed larger improvements in gender recognition with mismatched stimuli than the word training groups. Overall, the greatest improvement was seen for the mismatched stimuli in the gender training group. This result was expected given that the gender training group was explicitly trained to respond to voice gender during the training sessions.

One unexpected result is that the control group showed lower gender recognition accuracy for matched compared to mismatched stimuli. One possible explanation is that participants are interpreting the onset of the target voice as a new source with an associated change in gender (even when the gender is the same). This would result in lower accuracy for the matched condition. Training (of any of the three types) may reduce the tendency to hear the onset of the target voice as signaling a change in gender when no gender difference is present.

3. Daily training performance

Figure 4 summarizes the average training performance for the three training conditions. Word recognition performance, averaged across the 10 blocks and two daily training sessions, was 60% for the matched word training group and 64% for mismatched training. In contrast to the word recognition training groups, the gender training group approached ceiling rapidly on day 1.

The performance-adapted average TMR differed across the three training groups. The matched and mismatched training groups both had an average TMR of 5 dB for day 1 training across blocks 1–10. In comparison, the gender training group achieved a substantially lower TMR of –5 dB for blocks 1–10. There was little further improvement (decrease) in TMR for any of the three groups on day 2. Since the

FIG. 3. Boxplots (medians and interquartile ranges) of gender recognition difference scores (RAU) between PRE and POST tests (upper panel) and between LATE and PRE tests (lower panel) with matched gender target-masker combinations (dark gray bars) and mismatched combinations (light gray bars) for each training group. The “whiskers” represent 1.5 times the inter-quartile range; dots indicate outliers.

FIG. 4. Means and standard errors across listeners of daily training scores in each of the three training groups.

stimuli and number of trials were comparable across the training groups, the variation in TMR may have contributed to the differences in training outcomes reported in Sec. IV B.

V. GENERAL DISCUSSION

Experiment 1 confirmed findings in previous studies (e.g., Cullington and Zeng, 2008) that a difference in gender between target and masker voices can lead to significantly higher intelligibility of noise and tone-vocoded speech. Listeners showed higher word recognition accuracy for stimuli differing in voice gender (male target voice combined with female masker, or *vice versa*) compared to same-gender combinations. The amount of benefit increased with the number of channels and was independent of source type (noise or tone). These results motivated the selection of stimulus parameters for the training sessions in experiment 2, which explored the question of whether the benefits of voice gender differences could be enhanced by perceptual training. A significant improvement in word recognition accuracy was found in each of the three training conditions compared to a control condition with no training. Contrary to predictions, however, listeners who received word recognition training with sentence pairs differing in voice gender did not show greater improvement than those who received training with same-gender pairs. Training led to improved word recognition in all three experimental conditions, but its benefits were not selectively enhanced by voice gender differences. Listeners trained on a gender recognition task with the same stimuli showed comparable improvements.

The training effects reported here appear to be inconsistent with the findings of Massida *et al.* (2013) who examined the ability of CI users to discriminate male from female voices. Their stimuli were constructed using an auditory morphing technique to generate a continuum of sounds, ranging from a typical male to a typical female voice at the endpoints. Voice gender discrimination was poorer for CI users compared to listeners with normal hearing, and extremely variable near the midpoint of the continuum. They tested several groups of CI users with different durations of CI experience (post-activation delay), along with a set of ten CI users in a longitudinal study tested at comparable delays. In both cases they found that voice gender categorization did not improve substantially as a function of CI experience. One important difference is that Massida *et al.* tested CI users rather than normal-hearing listeners attending to CI simulations in the present study. However, there are several additional differences in the design that may account for the discrepancy. First, their stimuli were very brief (39.2 ms) time-compressed vocoded syllables. Our stimuli were pairs of competing sentences with natural prosody, syntax, and semantic predictability, and the task involved word recognition in the presence of speech maskers. Second, Massida *et al.* did not provide adaptive training or feedback as in the present study.

Several aspects of the results are consistent with Loebach *et al.* (2008) who compared the effects of perceptual training on a lexical task (word recognition) and two indexical tasks (talker and gender recognition) using an

eight-channel cochlear implant simulation. They showed that training tasks involving word transcription, talker identification and gender identification all yielded significant improvements, consistent with the present study. However, they found that training on talker identification and transcription produced greater improvements and more robust generalization compared to gender identification training. These authors suggested that training effectiveness depends on the degree of controlled attention, which was presumed to be lower for the two-alternative gender task, which showed performance close to ceiling at the start of training. A plausible reason for the difference in results is that the stimuli in our experiment included the added dimension of speech-on-speech masking, which was not present in the Loebach *et al.* (2008) study. The inclusion of a competing talker increases the level of task difficulty and the need for focused attention, which may account for the difference in results.

In contrast to Loebach *et al.*, gender recognition training in our study produced significant increases in experiment 2 for both the trained task (gender recognition) and word recognition despite near ceiling performance during the training blocks. Given that all training groups were presented the same number of trials, identical duration of training, and similar instructions to listen for the middle sentence despite differences in the training task, it is reasonable to conclude that the improvements were the result of active participation in training. Consistent with this conclusion is the observation that the performance of each of the training groups showed improvement on the respective task over time. The word recognition training groups began with performance lower than 50% correct on block 1, but improved to over 70% by the end of training on day 2, a pattern that was similar to the talker identification and sentence transcription training groups in Loebach *et al.* (2008). Taken together, these findings suggest that performance benefits may be the result of a rapid learning process that depends on active exposure to (and responding to) a challenging set of stimuli. However, similar improvements in each of the three training groups (and more specifically, the failure to observe greater improvement with mismatched-gender word recognition training) make it impossible to ascribe the effects to improved voice gender recognition or voice gender-based segregation.

Both of the experiments reported here indicate that voice gender differences can lead to significantly higher word recognition accuracy for masked speech processed by an eight-channel tone vocoder. Although some studies have indicated that CI users and listeners presented with eight-channel noise vocoder simulations receive little or at best inconsistent benefit from differences in voice gender between competing voices (Stickney *et al.*, 2004), more recent studies suggest that voice gender differences can lead to some degree of improvement in word recognition (e.g., Cullington and Zeng, 2008). The amount of improvement may be affected by differences in vocoder processing, the degree of predictability and redundancy of the speech materials, and characteristics of the talkers used as targets and maskers. Visram *et al.* (2012) showed that CI users can also benefit from voice gender differences in a similar manner to

the CI simulation listeners in the present study. Together with previous findings that voice gender discrimination is significantly correlated with speech recognition (Li and Fu, 2011), these results support the interpretation that voice gender can provide important cues for speech recognition for both CI users and vocoder simulations.

The eight-channel tone vocoder condition in experiment 1 and the pre-test condition of experiment 2 both resulted in word recognition scores that were higher for gender-mismatched compared to gender-matched stimuli (by 5.8 and 4.7 percentage points, respectively). However, the post-test condition of experiment 2 showed a mean improvement of 15.6 percentage points following training, suggesting that training can enhance the effects of voice gender differences. However, there was no difference between the training groups in the size of this effect: listeners who trained with gender-matched sentences showed the same amount of improvement as those who trained with gender-mismatched sentences. One interpretation is that perceptual learning of cues associated with voice gender differences is restricted or leads to limited gains for word recognition. Alternatively, it may also reflect a failure of voice gender training to generalize across stimuli (because training and test sessions used different speech materials and different talkers). Further research is necessary to distinguish these possibilities.

ACKNOWLEDGMENTS

Support for this research was provided by the National Institute of Deafness and other Communication Disorders F31 DC9537 to J.S. and National Science Foundation grant 1124479 to P.A. We thank our participants for their cooperation and support.

Alain, C., and Tremblay, K. (2007). "The role of event-related brain potentials in assessing central auditory processing," *J. Am. Acad. Audiol.* **18**(7), 573–589.

Assmann, P. F., and Katz, W. F. (2000). "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am.* **108**, 1856–1866.

Başkent, D., and Gaudrain, E. (2016). "Musician advantage for speech-on-speech perception," *J. Acoust. Soc. Am.* **139**, EL51–EL56.

Bierer, J. A., and Faulkner, K. F. (2010). "Identifying cochlear implant channels with poor electrode-neuron interface: Partial tripolar, single-channel thresholds and psychophysical tuning curves," *Ear Hear.* **31**(2), 247–258.

Brokx, J. P. L., and Nöteboom, S. G. (1982). "Intonation and the perception of simultaneous voices," *J. Phonetics* **10**, 23–26.

Brungart, D. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.* **109**, 1101–1109.

Chatterjee, M., and Peng S. C. (2008). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," *Hear. Res.* **235**, 143–156.

Cullington, H. E., and Zeng, F. G. (2008). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *J. Acoust. Soc. Am.* **123**, 450–461.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *J. Acoust. Soc. Am.* **114**, 2913–2922.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," *J. Exp. Psychol. Gen.* **134**(2), 222–241.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of

cochlear-implant signal processors with 6-20 channels," *J. Acoust. Soc. Am.* **104**, 3583–3585.

Fitch, W. T., and Giedd, J. (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Am.* **106**, 1511–1522.

Friesen, L. M., Shannon, R. V., Başkent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.* **110**, 1150–1163.

Fu, Q. J., Chinchilla, S., and Galvin, J. J. (2004a). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," *J. Assoc. Res. Otolaryngol.* **5**, 253–260.

Fu, Q. J., and Galvin, J. J. III (2007). "Perceptual learning and auditory training in cochlear implant recipients," *Trends Amplif.* **11**(3), 193–205.

Fu, Q. J., Galvin, J., Wang, X., and Nogaki, G. (2004b). "Effects of auditory training on adult cochlear implant patients: A preliminary report," *Cochlear Implants Int.* **5**, 84–90.

Fu, Q. J., Nogaki, G., and Galvin, J. J. III (2005). "Auditory training with spectrally shifted speech: Implications for cochlear implant patient auditory rehabilitation," *J. Assoc. Res. Otolaryngol.* **6**, 180–189.

Fu, Q. J., Shannon, R. V., and Wang, X. (1998). "Effects of noise and number of channels on vowel and consonant recognition: Acoustic and electric hearing," *J. Acoust. Soc. Am.* **104**, 3586–3596.

Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q. J., Free, R. H., and Başkent, D. (2014). "Gender categorization is abnormal in cochlear implant users," *J. Assoc. Res. Otolaryngol.* **15**, 1037–1048.

Gaudrain, E., and Başkent, D. (2015). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," *J. Acoust. Soc. Am.* **137**, 1298–1308.

Geers, A., Brenner, C., Nicholas, J., Uchanski, R., Tye-Murray, N., and Tobey, E. (2002). "Rehabilitation factors contributing to implant benefit in children," *Ann. Otol. Rhinol. Laryngol.* **111**, 127–130.

Gifford, R. H., and Revit, L. J. (2010). "Speech perception for adult cochlear implant recipients in a realistic background noise: Effectiveness of preprocessing strategies and external options for improving speech recognition in noise," *J. Am. Acad. Audiol.* **21**(7), 441–451.

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

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

Hillenbrand, J. M., and Clark, M. J. (2009). "The role of f0 and formant frequencies in distinguishing the voices of men and women," *Atten. Percept. Psychophys.* **71**, 1150–1166.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.* **27**, 187–207.

Krull, V., Luo, X., and Iler Kirk, K. (2012). "Talker-identification training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition," *J. Acoust. Soc. Am.* **131**, 3069–3078.

Li, T., and Fu, Q. J. (2011). "Voice gender discrimination provides a measure of more than pitch-related perception in cochlear implant users," *Int. J. Audiol.* **50**, 498–502.

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.

Loebach, J. L., and Pisoni, D. B. (2008). "Perceptual learning of spectrally degraded speech and environmental sounds," *J. Acoust. Soc. Am.* **123**(2), 1126–1139.

Loizou, P. C., Dorman, M. F., Tu, Z., and Fitzke, J. (2000). "Recognition of sentences in noise by normal-hearing listeners using simulations of speech-type cochlear implant signal processors," *Ann. Otol. Rhinol. Laryngol.* **185**, 67–68.

Massida, Z., Marx, M., Belin, P., James, C., Fraysse, B., Barone, P., and Deguine, O. (2013). "Gender categorization in cochlear implant users," *J. Speech Lang. Hear. Res.* **56**, 1389–1401.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.* **95**(2), 1085–1099.

Nogaki, G., Fu, Q. J., and Galvin, J. J. III (2007). "Effect of training rate on recognition of spectrally shifted speech," *Ear Hear.* **28**(2), 132–140.

- Qin, M. K., and Oxenham, A. J. (2005). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," *Ear Hear.* **26**, 451–460.
- Schvartz, K. C., and Chatterjee, M. (2012). "Gender identification in younger and older adults: Use of spectral and temporal cues in noise-vocoded speech," *Ear Hear.* **33**, 411–420.
- Skinner, M. W. (2003). "Optimizing cochlear implant speech performance," *Ann. Otol. Rhinol. Laryngol.* **112**, 4–13.
- Skuk, V. G., and Schweinberger, S. R. (2014). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," *J. Speech Lang. Hear. Res.* **57**, 285–296.
- Smith, D. R., and Patterson, R. D. (2005). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *J. Acoust. Soc. Am.* **118**, 3177–3186.
- Stickney, G. S., Zeng, F. G., Litovsky, R., and Assmann, P. F. (2004). "Cochlear implant speech recognition with speech maskers," *J. Acoust. Soc. Am.* **116**, 1081–1091.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Sullivan, J. R., Thibodeau, L., and Assmann, P. F. (2013). "Auditory training in interrupted noise and improvements in speech recognition in noise for children with hearing impairment," *J. Acoust. Soc. Am.* **133**(1), 495–501.
- Tobey, E. A., Geers, A. E., Brenner, C., Altuna, D., and Gabbert, G. (2003). "Factors associated with development of speech production skills in children implanted by age five," *Ear Hear.* **24**, 36S–45S.
- Tremblay, K. L. (2007). "Training-related changes in the brain: Evidence from human auditory-evoked potentials," *Semin. Hear.* **28**(2), 120–137.
- Visram, A. S., Kluk, K., and McKay, C. M. (2012). "Voice gender differences and separation of simultaneous talkers in cochlear implant users with residual hearing," *J. Acoust. Soc. Am.* **132**(2), EL135–141.
- Zeng, F.-G., Tang, Q., and Lu, T. (2014). "Abnormal pitch perception produced by cochlear implant stimulation," *PLoS One* **13**, e88662.