JSLHR

Research Article

# Auditory Learning Using a Portable Real-Time Vocoder: Preliminary Findings

Elizabeth D. Casserly[a] and David B. Pisoni[a]

**Purpose:** Although traditional study of auditory training has been in controlled laboratory settings, interest has been increasing in more interactive options. The authors examine whether such interactive training can result in short-term perceptual learning, and the range of perceptual skills it impacts.
**Method:** Experiments 1 ($N$ = 37) and 2 ($N$ = 21) used pre- and posttest measures of speech and nonspeech recognition to find evidence of learning (within subject) and to compare the effects of 3 kinds of training (between subject) on the perceptual abilities of adults with normal hearing listening to simulations of cochlear implant processing. Subjects were given interactive, standard lab-based, or control training experience for 1 hr between the pre- and posttest tasks (unique sets across Experiments 1 & 2).
**Results:** Subjects receiving interactive training showed significant learning on sentence recognition in quiet task (Experiment 1), outperforming controls but not lab-trained subjects following training. Training groups did not differ significantly on any other task, even those directly involved in the interactive training experience.
**Conclusions:** Interactive training has the potential to produce learning in 1 domain (sentence recognition in quiet), but the particulars of the present training method (short duration, high complexity) may have limited benefits to this single criterion task.

The ability of listeners with normal hearing (NH) to perceive speech and other acoustic signals through simulations of cochlear implant (CI) processing has been studied extensively (e.g., Bent, Buchwald, & Pisoni, 2009; Carroll & Zeng, 2007; Fu, Shannon, & Wang, 1998; Loebach & Pisoni, 2008; Loebach, Pisoni, & Svirsky, 2010; Shannon, Zeng, Kamath, & Ekelid, 1995). Using subjects with NH under CI simulation as a model system for CI users who are deaf provides many research benefits, from increasing feasible sample size to effectively isolating the effects of peripheral, signal-processing implant factors on accuracy limits in the perception of speech and other sounds. Due to their accessibility and perceptual robustness, populations with NH are also ideal for initial exploration of novel signal-processing techniques and perceptual training protocols (e.g., Baskent & Shannon, 2003; Loebach et al., 2010). Although any research conducted with NH models must then be connected back to the deaf CI-user population, these investigations help ensure that later work investigating CI users is focused and likely to succeed in producing substantive results.

Among the many current research questions revolving around the experience of CI users who are deaf, one of the foremost concerns possible avenues of intervention or training for individuals who are at high risk for poor outcomes with their devices. Some CI users receive tremendous benefits from their devices and are able to use spoken communication with a high level of fluency and flexibility in the quiet; others, however, enjoy only limited gains from their CIs and are often unable to discriminate much more than the presence versus absence of sound in their environment (Dorman, Dankowski, McCandless, Parkin, & Smith, 1991; Geers, Brenner, & Davidson, 2003; Geers, Tobey, & Moog, 2011; Kaiser, Kirk, Lachs, & Pisoni, 2003; Pisoni, Svirsky, Kirk, & Miyamoto, 1997). Understanding the underlying causes of such outcome variation is paramount; until we know what factors are at play, training and treatments intended to help those CI users on the lower end of the distribution can be only ad hoc, at best.

Perhaps equally important, however, is maximizing the efficacy and everyday utility of any training given to individuals needing help, both in terms of efficiency and true outcome improvements. Research using subjects with NH is well suited for addressing these issues, enabling speech scientists to explore the possible cost–benefit gains for a wide variety of training methods. Using work with NH models of CI perception, for example, has shown that a variety of training materials can be effective in promoting perceptual learning, from isolated words to connected speech to environmental

[a]Speech Research Laboratory, Indiana University, Bloomington

Correspondence to Elizabeth D. Casserly, who is now at Trinity College, Hartford, CT: elizabeth.casserly@trincoll.edu

Editor: Nancy Tye-Murray
Associate Editor: David Moore

Received August 12, 2013
Revision received February 14, 2014
Accepted January 15, 2015
DOI: 10.1044/2015_JSLHR-H-13-0216

*Disclosure: The authors have declared that no competing interests existed at the time of publication.*

sounds (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Fu, Nogaki, & Gavin, 2005; Loebach & Pisoni, 2008; Loebach et al., 2010; Shafiro, 2008a, 2008b; Shafiro, Sheft, Gygi, & Ho, 2012). We also know that providing learners with immediate feedback is critical (Davis et al., 2005) and that such feedback can be text based, and therefore available for use in training with CI users who are deaf (Loebach et al., 2010).

The standard auditory training model emerging from such studies is one in which subjects hear a brief sample of speech or other environmental sound, attempt to recognize or transcribe what they have heard, and are given the sound again following their recognition attempt, this time with the correct transcription or identification shown simultaneously on a computer screen. Relatively brief periods of training using this method (approximately 1 hr) result in improvements to perceptual recognition of similar stimuli (e.g., similar, but novel, speech or environmental sound materials), and frequently to improved recognition of new, untrained stimulus categories, such as isolated words, following training on stimuli at the sentence level (Fu et al., 2005; Loebach, Bent, & Pisoni, 2008; Loebach & Pisoni, 2008; Loebach et al., 2010; Shafiro et al., 2012). Such generalization from training materials to novel acoustic stimuli is notoriously challenging to achieve in studies of perceptual learning, but success is critical for a protocol to have any meaningful impact on the communicative outcome of a CI user.

There is a disconnect, however, between this type of conventional auditory training and the needs and challenges of real-world CI users. First, the central role of a stationary computer and prerecorded stimulus set limits both the appeal of the training for users and its adaptability to their unique needs or degree of experience with standard speech assessment materials. Second, there is reason to believe that improvements in standard measures of speech recognition—such as sentence or word transcription in the quiet—may not coincide with user-perceived listening benefits (cf. Henshaw & Ferguson, 2013). That is, despite the gains seen in the measured tasks, listeners with a hearing impairment report continuing difficulty with other activities, such as listening in noise or understanding accented speech. Although adjustments to the standard training may allow it to promote learning that generalizes to these areas, such generalization across tasks appears so far to be limited in scope (Loebach, Pisoni, & Svirksy, 2009; but see Ferguson, Henshaw, Clark, & Moore, 2014).

A method of learning that does not rely on predetermined locations and stimuli, which could be used to expose listeners to a variety of real-world tasks, might therefore be useful in augmenting our current notion of optimal auditory training. A device was developed recently that makes preliminary exploration of such a learning method possible. The portable real-time vocoder (PRTV; Casserly, Pisoni, Smalt, & Talavage, 2011) performs continuous CI simulation of all ambient acoustics, playing the resulting noise-vocoded signals with very little delay (<10 ms) over noise-isolating insert earphones (see details in Method). It is crucial to note that the acoustic transformation is performed using an iPod (Apple, Cupertino, CA), making it extremely lightweight and portable. As a result, CI simulation of acoustic signals can be conducted virtually anywhere, on acoustic signals varying from city buses to a listener's own voice, and listeners can experience vocoded acoustics during a variety of tasks and under a variety of circumstances.

In this article, we report the results of two preliminary experiments exploring the effects of this type of interactive, high-variability exposure or "training" with vocoded acoustic signals. Each experiment gave listeners 1 hr of experience listening and talking through the PRTV and assessed their ability to perform a range of perceptual tasks following training. In all cases, performance of these *interactive* subjects was compared with that of subjects who were given the standard laboratory training described above and with that of a control group who were given no training with vocoded acoustics. Our goals in these experiments were twofold: First, to determine whether naturalistic, interactive training can result in perceptual learning of CI-simulated signals, and second, to examine performance on a diverse set of perceptual tasks to discover areas where the results of interactive training may diverge from the outcomes of standard auditory training.

## Experiment 1

Allowing subjects with NH to experience real-time CI-simulated acoustics through a portable, interactive system introduces a number of characteristics to their training exposure that standard methods generally lack. Listening can occur in multiple environments, for example, including areas with mechanical and multitalker background noise, which gives subjects experience with recognizing speech in noise, as well as exposure to a variety of nonspeech environmental sounds. Subjects can also experience face-to-face communication, giving them direct opportunities for audiovisual (AV) integration and the use of lip reading to aid in perceiving vocoded speech. They can engage in connected, meaningful conversation with a social partner, motivating them to use and successfully recognize aspects of speech such as prosody and pragmatic meaning (e.g., sarcasm, indirect requests, etc.).

Each of these novel facets of interactive training represents an additional challenge for the subject but also a domain of potential improvement for their perceptual skills. Experiment 1, therefore, was designed to assess not only whether subjects given interactive training were able to show perceptual learning at all but also whether these subjects showed any gains in areas related to novel training characteristics, compared with either subjects given standard laboratory training or those in the control, no-training condition.

Learning as a result of training was measured within each group of subjects (interactive, laboratory, control) using a pretest–posttest comparison for performance on a task requiring recognition of CI-simulated sentence-level speech materials presented in the quiet. Subjects given the standard lab-based training were expected to show significant improvements on this task, whereas control subjects were expected to show no gains. Interactive subjects, despite the unique challenges of their training exposure, were also expected to show significant improvements in this basic speech perception task.

Following the speech-in-quiet posttest measure, subjects were given several additional tasks designed to tap into the differences between interactive and standard training mentioned above. All three groups were asked to perform sentence recognition in noise, isolated word recognition, audiovisual word recognition, discrimination of prosodic contours, and environmental sound recognition. Speech recognition in noise, audiovisual word recognition, prosodic contour discrimination, and environmental sound recognition are all tasks that interactive subjects encountered during their multienvironment, conversational training (see Method for details); it was therefore expected that each of these processing domains might be areas where they would show indirect evidence of perceptual learning by outperforming lab-trained subjects or controls. Audio-only isolated word recognition, by contrast, was not a domain in which interactive subjects were expected to show particular, differential improvement, but it was included both as a task where lab-trained subjects might show generalization from their sentence-based training (see Loebach & Pisoni, 2008) and also as a means of calculating subjects' *audiovisual benefit*, a standard assessment of AV perceptual gains that takes an individual's audio-only perceptual skill into account when interpreting their AV recognition accuracy (Sumby & Pollack, 1954).

In total, therefore, subjects in all three training groups completed one pretest task (sentence recognition in quiet), 1 hr of training, and six posttest tasks designed to assess gains in perceptual skill across a number of processing domains. Pretest versus posttest improvements were analyzed to show learning as a result of training, whereas comparisons between the interactive and lab subjects versus the controls were conducted to test improvements indirectly. Repetition of every perceptual task as both a pretest and a posttest was not included to limit the opportunities for implicit learning during the pretest phase of the experiment and to avoid biasing interactive and lab-trained subjects, alerting them toward particular aspects of the acoustic signal that would be tested following training.

## Method

All of the methods were approved for use with human subjects by the Indiana University Institutional Review Board.

*Subjects.* A total of 37 subjects (13 men, 24 women) were recruited from Indiana University and the surrounding community to participate in this experiment. All subjects were healthy, young, adults between the ages of 18 and 30 years who were monolingual speakers of North American English; four of the volunteers were excluded from data analysis for failing to meet the native language criterion, leaving 33 remaining. Before participating, all subjects passed an audiometric screening assessment to ensure that their hearing thresholds were ≤ 25 dB between 500 and 8000 Hz, equivalent to NH acuity (<25 dB hearing loss [HL]).

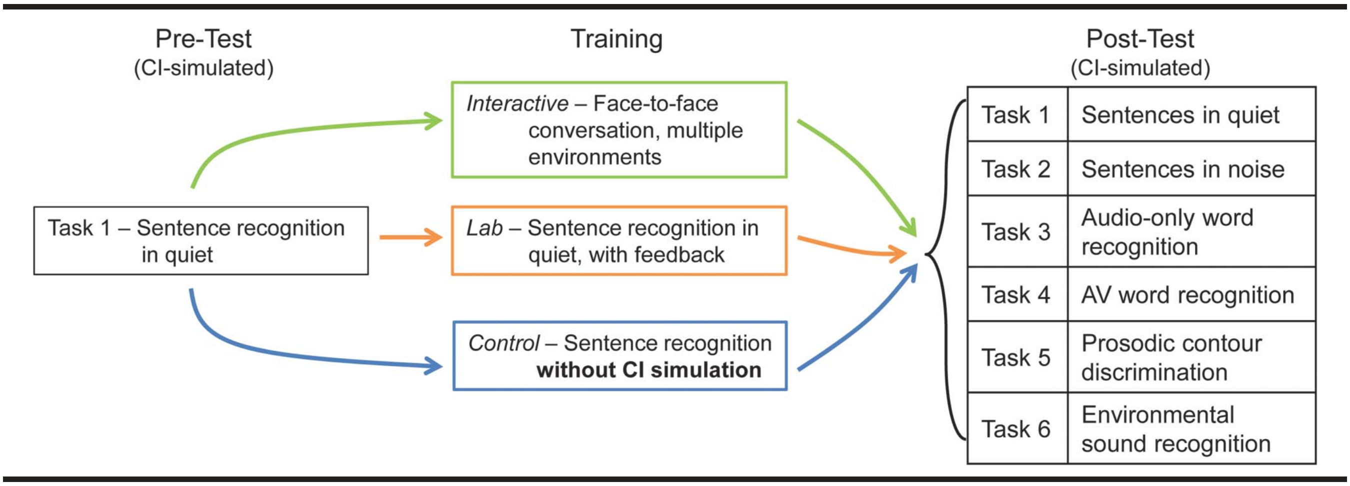
*Acoustic transformation.* Simulation of cochlear implant processing was carried out using a PRTV, a lightweight device consisting of a compact, high-speed processor (8 GB iPod touch, model A1367), input microphone (Williams MIC090 mini lapel clip; Williams Sound, Eden Prairie, MN), output noise-isolating earphones with disposable single-use foam tips (Etymotic HF5; Etymotic, Elk Grove Village, IL), and a set of noise-attenuating headgear (Elvex SuperSonic; Elvex, Bethel, CT). Custom signal-processing software took continual sound input from the lapel microphone, worn on the attenuators just above a subject's left ear, applied the acoustic transformation (cf. Kaiser & Svirsky, 1999, 2000), and relayed the modified acoustic signal to the insert earphones with less than a 10-ms delay (Casserly et al., 2011).

Sampled acoustics were subjected to a broad bandpass frequency filter of 252 to 7000 Hz then split into eight nonoverlapping frequency bands. The amplitude within each of these channels was then used to generate envelope-matched noise of the same bandwidths, which were summed to create a noise-vocoded acoustic simulation of a CI. Resolution within each channel was lost, therefore, during the transformation, along with any information occurring at frequencies higher than 7000 Hz or lower than 252 Hz. Subjects completing the interactive auditory training, however, received additional perceptual feedback during their own speech production in the form of bone conduction (cf. Stenfelt & Håkansson, 2002). This feedback was not digitally countered or masked.

*Experimental design.* Subjects were randomly assigned to one of three groups: Those receiving interactive exposure (interactive), those completing standard laboratory-based training (lab), and those completing the lab-based training protocol without the vocoding transformation (control). All subjects completed identical pretest and posttest assessments under conditions of real-time CI simulation; only their experiences and activities during a 1-hr learning period between tests differed across groups. The structure of the pretest, training period, and posttest in Experiment 1 are summarized in Figure 1.

During the 1-hr exposure period, subjects in the interactive group listened to CI-simulated acoustic signals while engaging in naturalistic, active-hearing, communicative and environmental interactions. Interactive subjects conversed with the researcher administering the experiment (either the first author or an experienced research assistant) and moved about the area near the testing location. These subjects therefore experienced vocoded acoustics in a wide range of conditions, including reverberant spaces (e.g., empty hallways); quiet, ideal listening conditions (testing room); outdoors, near traffic and in relatively quiet courtyards; in the presence of white noise (near ventilation/air conditioning units); and surrounded by competing multitalker babble (e.g., student lounge). During spoken interaction in each of these locations, the experimenter encouraged each subject to speak frequently and to test their understanding of her words to them. Such explicit discussion of recognition accuracy was the only feedback provided, other than the successful flow of communication between the experimenter and subject.

**Figure 1.** Schematic of the design of Experiment 1. All subjects completed the same pre- and posttests following training; only the type of training received by each subject varied (interactive, lab, control). CI = cochlear implant.


Pre-Test
(CI-simulated)
Task 1 – Sentence recognition in quiet
Training
*Interactive* – Face-to-face conversation, multiple environments
*Lab* – Sentence recognition in quiet, with feedback
*Control* – Sentence recognition **without CI simulation**
Post-Test
(CI-simulated)
Task 1 Sentences in quiet
Task 2 Sentences in noise
Task 3 Audio-only word recognition
Task 4 AV word recognition
Task 5 Prosodic contour discrimination
Task 6 Environmental sound recognition

By contrast, subjects in the lab exposure group spent the 1-hr exposure period completing a relatively passive computer-based training protocol that has been shown to be highly effective in promoting learning for acoustic perception of CI-simulated signals (e.g., Loebach et al., 2010). Subjects were asked to transcribe meaningful English sentences (see Stimulus Materials) heard through the PRTV, and after entering each response, immediate feedback was given in the form of the correct orthographic transcription coupled with a second repetition of the acoustic stimulus. Training was timed to have a duration of 1 hr; the total number of training stimuli experienced by each subject was therefore variable across subjects.

Because of the natural dynamics of conversation, the amount and nature of acoustic input to subjects in the interactive group also varied across sessions but more substantively than for the lab exposure subjects. This asymmetry is a potential confounding factor in the current investigation, but its inclusion was deliberate: Such imbalance is an inevitable consequence of the nature of the two exposure types and therefore is not undesirable in an initial comparison of the effects of interactive versus lab-based perceptual learning.

In contrast to the interactive and lab subjects, subjects in the third (control) group did not experience any transformation of acoustic signals during their 1-hr exposure period. They completed the same sentence-transcription training given to lab subjects but with the PRTV transmitting unprocessed signals.

Following the training period, subjects in all three groups completed the same set of six tasks to assess their perceptual performance through the CI simulation. The six tasks were as follows: (a) sentence transcription in quiet, (b) sentence transcription in the presence of multitalker babble, (c) open-set word recognition in the quiet, (d) AV open-set word recognition in the quiet, (e) prosodic contour identification (in quiet), and (f) environmental sound recognition (in quiet). No feedback was given on any of these tasks, although two practice trials were given, with feedback, at the start of Task 2 (sentences in babble), Task 5 (prosodic contour identification), and Task 6 (environmental sound recognition), to orient subjects to these unfamiliar perceptual processing tasks.

## Materials and Stimuli

*Training materials for lab and control groups.* A set of 135 meaningful English sentences (75 Harvard/Institute of Electrical and Electronics Engineers [IEEE] sentences [1969] and 60 Boys Town sentences [Stelmachowicz, Hoover, Lewis, Kortekaas, & Pittman, 2000]) was used for this sentence transcription in quiet task. None of the sentences occurred in any other portion of the experiment. Sentences from the IEEE database contained five keywords and varied syntactic structure (e.g., *The two met while playing on the sand*), whereas those from the Boys Town materials were syntactically simple phrases of four total words each (e.g., *Some birds eat worms*). All of the training sentences were produced by a single female talker from the Midland dialect region and digitally edited to have the same average root-mean-square decibel sound level, which was also used for the target materials in Tasks 1 to 6.

*Task 1: sentence transcription (in quiet).* For the sentence transcription in quiet task, the only test completed both pre- and posttraining, 20 sentences were randomly selected from lists 1 to 4 of the Harvard/IEEE sentence database, and 10 were randomly assigned to the pre- and posttest assessments. Accuracy on the sentence transcription task was measured by scoring the number of keywords correctly recognized by each subject (including obvious typographical errors as correct responses). Each sentence contained five keywords, resulting in a total of 50 keywords per test set. The sentences used in this task were recorded by a second female talker from the Midland dialect region (distinct from the talker in the training materials) and leveled to the same average sound pressure level (SPL). This task, along with the other five tasks described below, was completed while subjects experienced real-time CI acoustic simulation through the PRTV.

*Task 2: sentence transcription (in babble).* For this sentence-recognition-in-babble task, 15 sentences were randomly selected from lists 1 to 4 of the Hearing in Noise Test (HINT; Nilsson, Soli, & Sullivan, 1994) to be used as stimuli. These sentences were shorter, contained fewer keywords (between three and five), and had simpler syntactic structure than the materials in the IEEE database (e.g., *They like orange marmalade*). HINT sentences are used extensively in audiological testing, and the standard set of recordings used in the clinic was also used here. These sentences therefore contained speech from a third distinct talker, a male speaker of a relatively unmarked or standard dialect of American English.

HINT stimuli for Task 2 were mixed at a signal-to-noise ratio (SNR) of +8 dB with four-talker babble generated from the speech of two male and two female talkers. The babble track was generated for the purposes of this experiment by concatenating sentences from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993) spoken by randomly selected talkers of the target genders from the South Midland dialect region (local to the testing location). The sentences were digitally manipulated to have the same average SPL across items and talkers, 8 dB below that of the HINT sentences. One hundred sentences (approximately 4 min of speech) from each talker were then concatenated randomly and summed across talkers so that no two sentences began or ended simultaneously and no periods of silence greater than 100 ms occurred for any given talker. The clip of babble used for each HINT sentence stimulus was selected by randomly excising a time-matched portion of the merged four-talker speech file and combining it with the target sentence.

*Task 3: Open-set word recognition.* Forty items were selected from the Lexical Neighborhood Test (LNT; Kirk, Pisoni, & Osberger, 1995) database of monosyllabic, isolated English words for use in this task. The LNT database contains only items that are familiar to undergraduate students on the basis of ratings from Nusbaum, Pisoni, and Davis (1984) and that have been designated as either “easy” or “hard” to recognize based on the structure of their lexical neighborhood. The more phonetically similar phonological neighbors a word has and the higher the frequency of those neighbors, the more potent competition it must overcome to be recognized successfully, making recognition slower and less accurate overall (Luce & Pisoni, 1998). Words were selected semirandomly from the LNT database such that 20 words of each difficulty were included. Each set of 20 easy/hard items was also balanced for its phonetic content, so the lists contained equal numbers of high versus low and front versus back vowels. These balancing efforts were made to ensure that our total set of isolated words covered a broad range of phonetic and lexical content.

Recorded versions of the LNT stimuli were then taken from the Hoosier Audiovisual Multitalker database (Sheffert, Lachs, & Hernandez, 1996). This set of materials contains audiovisual productions of LNT items from a number of male and female talkers; utterances from a single female speaker, distinct from the speakers in Tasks 1 and 2, were selected for use in Experiment 1. Audio files were extracted from the AV tokens to generate the auditory-only stimuli for this task, which were heard by all subjects through the PRTV.

*Task 4: Audiovisual open-set word recognition.* An additional, nonoverlapping set of 40 monosyllabic words from the LNT database was selected for use in the audiovisual word recognition task. The same selection criteria described for Task 3 were used in constructing this set of items as well, including balancing neighborhood effects and vowel quality content. Tokens produced by the same talker used in Task 3 were used for this task as well, but in this condition the full synchronized AV recordings of each word in the Hoosier Audiovisual Multitalker database were used. Each AV signal displayed the face and upper shoulders of the female talker, wearing a black turtleneck, in front of a gray background. All of the words were produced with a neutral facial expression, and the speaker’s gaze was focused directly at the camera. Background noise and average SPL in Task 4 stimuli were identical to those word tokens used in Task 3. Using scores from these two tasks, therefore, allowed us to calculate the perceptual gains that each subject received from the presence of AV, as opposed to auditory-only, speech information.

*Task 5: Prosodic contour identification (in quiet).* To assess listeners’ ability to perceive meaningful pitch-based contrasts in vocoded speech materials, we asked subjects to label the type of intonation used in tokens of isolated spoken words. A database containing sets of 25 words spoken with either declarative (falling f0) or interrogative (Initial Fall + Final Rise in f0) prosodic contours (Casserly, 2013) was used to provide the stimuli for this task. Each word contained only sonorant segments (vowels, glides, liquids, and nasals), allowing the f0 information in the acoustic signal to be uninterrupted.

Twenty items each were selected from the productions of two talkers, one man and one woman. Utterances with artifacts such as background noise or speech errors (five per talker) were eliminated. Talkers were chosen on the basis of their f0 production: They had the smallest and largest contrasts, respectively, between statement and question intonations of the 12 talkers available in the database. The contrast between these prosodic contours is greatest at the end of an utterance, where the final fall in f0 associated with declarative statements is maximally different from the final rise in f0 associated with questions in American English. For these two speakers, the average difference in pitch between statement tokens and question tokens utterance finally was 327.29 Hz for the female speaker and 70.22 Hz for the male speaker. For reference, average vocal pitch for these speakers’ tokens was 290.14 Hz and 129.29 Hz, respectively.

Each talker’s set of 20 items was randomly split between question and statement intonations so that 10 of each type of trial appeared in each iteration of the task. Trials were blocked by talker, and the order of presentation was counterbalanced across subjects. Within each single-talker block, item order was randomized. This task was also completed while all subjects were wearing an actively vocoding PRTV.

*Task 6: Environmental sound recognition (in quiet).* A set of 55 environmental sound stimuli was selected for use in this recognition task (Loebach & Pisoni, 2008; Shafiro, 2008a). A wide range of attested sound sources were represented, from human nonspeech vocalizations (e.g., laughing, infant crying) to transportation sounds (e.g., motorcycle engine, helicopter) to natural phenomena (e.g., rain, thunder). Subjects heard these sounds through the PRTV, without context, and were asked to identify them in an open-set naming task. Two practice trials containing different categories of sounds (elephant vocalization, zipper) were given, with feedback, to orient subjects to the task and the range of possible environmental sound sources included.

## Procedure

Subjects were randomly assigned to one of three training groups: Interactive, lab, or control. Before beginning the experiment, all subjects' hearing thresholds were tested using a portable audiometer (Ambco 650A), to ensure that their hearing was within normal ranges (<25 dB HL). Each subject was then seated in an Industrial Acoustics Corp. sound-attenuating booth in front of a computer screen and keyboard. Stimuli in the pretest and posttest tasks were presented via a loudspeaker calibrated to an average SPL of 70 dB across utterances that was held constant for each listener. These tasks, identical for subjects in all three groups, took between 5 and 10 min each to complete, for a total of approximately 45 min spent outside of training.

Between the pre- and posttest tasks, subjects experienced a period of 1 hr with the PRTV to adapt to the novel perceptual transformation of CI simulation. Each exposure period (see Experimental Design for details) was timed to last precisely 1 hr, regardless of the amount of conversation (interactive group) or the number of sentence stimuli (lab and control groups) completed in that time.

Subjects responded to all computer-based training trials and five of six nontraining tasks using a keyboard. Speech recognition tests (i.e., Tasks 1, 2, 3, and 4) required listeners to transcribe each utterance as accurately as possible; transcriptions of the AV stimuli were made via pencil and paper, rather than the computer keyboard, for the word recognition in Task 4. The AV stimuli were presented on the same computer display used for subjects' response entry, but text input was disabled for these trials. In Task 5 (prosodic contour identification), subjects were told that each stimulus was produced as either a statement or a question, and they were to indicate via button-press which of the two options they detected on each trial. For Task 6 (environmental sound recognition), subjects were asked to complete an open-set identification task, entering a description of the source of each sound they heard through the PRTV, being as specific as possible. In all six tasks, guessing was encouraged.

## Data Analysis

Listener responses to each of the six pretest–posttest tasks were converted to accuracy scores for analysis. Sentence recognition responses (Tasks 1 and 2) were scored by keyword accuracy; if a subject correctly transcribed any of the target keywords, regardless of serial position, they were given credit for recognizing it. Half credit was awarded for any responses that deviated from keywords by one phonetic segment (e.g., *boy* or *Troy* for target *toy*) or in morphological marking of number or tense (e.g., *plays* or *playing* for target *played*). Isolated word recognition tests (Tasks 3 and 4) were assessed using binary scoring: Subjects' responses either matched the target word exactly (barring homophone spelling substitutions; e.g., *pail/pale*) or they were counted as incorrect. Accuracy in prosodic contour identification was similarly straightforward, with subjects either correctly identifying the intonation of each utterance or not. Accuracy in environmental sound recognition was calculated by awarding full credit for correct responses or their synonyms (e.g., *chickens*, *clucking*, or *hens* would all be correct responses to the sound of chickens clucking), whereas half credit was awarded in the case of scope mismatches (as in a response of *music* for a strumming banjo).

Each subject's average accuracy scores in the pretest and six posttests were then used in a series of analyses to assess learning and generalization. First, initial group equivalency was calculated on the pretraining sentence recognition task. Second, the accuracy of subjects within each group was compared in the pre- and posttest sentence recognition task to directly assess the degree of perceptual learning. Third, performance across groups was compared for all six of the posttest assessments, to determine whether learning was achieved in any of these additional perceptual skills.

## Results

*Initial group equivalency.* Because subjects were randomly assigned to three independent groups, the initial equivalence of the groups' average perceptual abilities needed to be established. A one-way analysis of variance (ANOVA) with group as the factor of interest was therefore conducted first on pretest sentence recognition accuracy scores. As shown in Figure 2, average accuracy was very similar across the three groups, ranging from 53.7% (lab group) to 61.2% (control), and the ANOVA did not return a significant effect of group, $F(2, 30) = 0.930$, $p = .406$. Thus, subjects' initial abilities to perceive speech through the CI simulation did not differ significantly among groups.

*Within-group evidence of perceptual learning.* To assess the learning achieved by each group during the exposure period, a repeated-measures analysis of variance (RM-ANOVA) was conducted on the sentence recognition accuracy scores from the pre- and posttraining tests.[1] This RM-ANOVA returned a significant main effect of Learning (comparison of pretest vs. posttest accuracy), $F(1, 30) = 34.663$, $p < .00001$,

[1] The sphericity assumption of this RM-ANOVA was met; the equality of variance (homoscedasticity) assumption was met for the pretest data, but not for the posttest results. Correction for this violation has been applied in the *p* values reported in this section.

**Figure 2.** Pretest Task 1 (sentence recognition in quiet) performance by training group. Here and elsewhere, boxplots show median accuracy scores for each group (solid horizontal line) within shaded 25th to 75th percentile ranges. Error bars reflect 95% confidence intervals, and outliers are shown as individual points. Differences between groups were not significant prior to training.

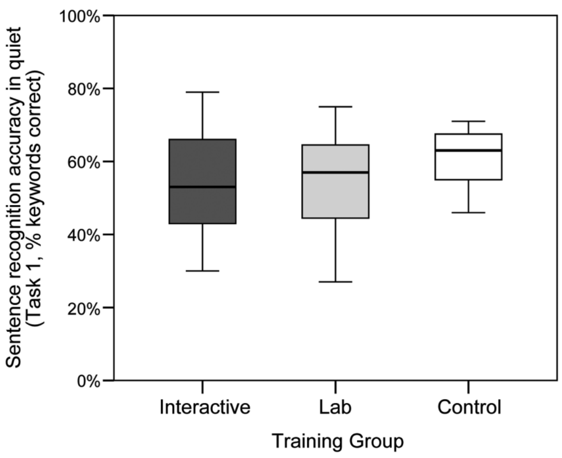
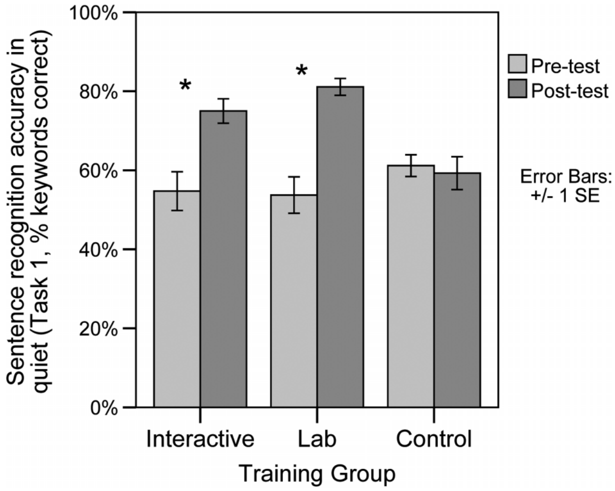


**Figure 3.** Mean pre- and posttest performance on Task 1 (sentence recognition in quiet) across the interactive, lab, and control groups. Subjects in the interactive and lab training groups improved significantly from pre- to posttest; control subjects did not. Error bars = ± 1 standard error (SE).



$\eta_p^2 = .536$, along with a significant Learning × Group interaction, $F(2, 30) = 11.597$, $p < .001$, $\eta_p^2 = .436$, but no main effect of group, $F(2, 30) = 1.471$, $p = .246$, $\eta_p^2 = .089$.

Simple paired-samples *t* tests, corrected for multiple comparisons ($\alpha = .0167$), were then conducted on the data from each group to examine the source of this significant interaction. These results are summarized in Figure 3. Average accuracy for interactive subjects went from 54.7% before training to 75.0% afterward, a significant improvement, $t(10) = 5.037$, $p = .001$. Subjects in the lab-trained group also showed significant effects of learning, with average recognition accuracy in quiet rising from 53.7% at pretest to 81.1% at posttest, $t(10) = 6.232$, $p < .001$. By contrast, subjects in the control group, who did not receive additional experience with the CI simulation during the 1-hr period between pre- and posttest assessments, did not show any significant changes in performance, $t(10) = -0.383$, $p = .710$, with pretest accuracy of 61.2% being followed by an accuracy of only 59.3% at posttest. Improvements in speech recognition, therefore, were obtained as a result of both interactive and standard lab exposure, whereas subjects in the control group, who were not given any experience with CI simulated acoustics outside of the pre- and posttests, did not show any evidence of learning.

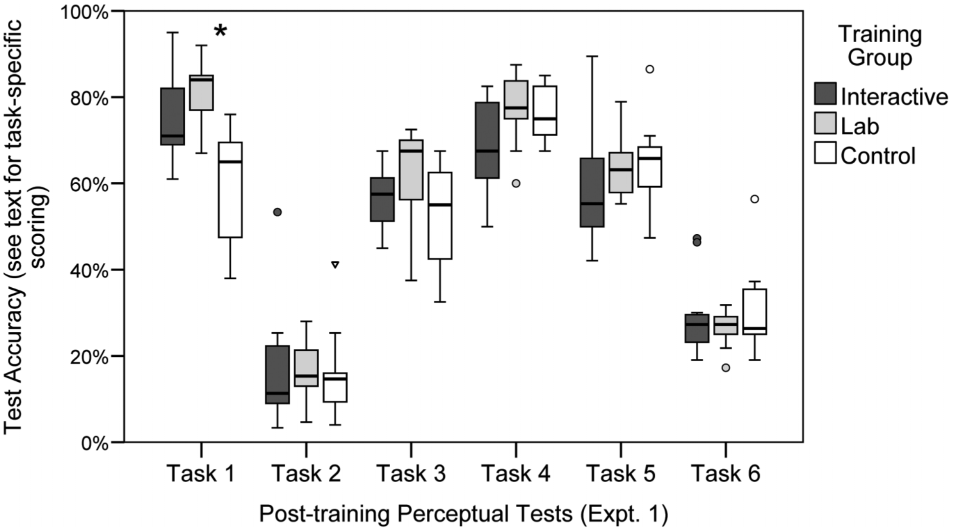
*Comparison across training types.* Differences among training groups (interactive, lab, and control) were also assessed in each of the six posttest tasks and in the derived AV benefit scores using ANOVA with Bonferroni correction for multiple comparisons ($\alpha = .007$, correction for seven simultaneous comparisons). The data are summarized in Figure 4. Levene's test of equality of variance was conducted on each ANOVA, and the data from each task except Task 1 was found to meet the homoscedasticity assumption across groups. The violation of homoscedasticity for Task 1 (visible in Figure 4 as a reduced spread in the accuracy scores of lab subjects relative to controls) is a potential issue for interpretation of the ANOVA results; however, having equal sample size across groups minimizes the effects of this violation and should not result in a greater risk of Type I error (Glass, Peckham, & Sanders, 1972).

A corrected one-way ANOVA on each of the six posttest tasks returned only one significant difference among groups: A significant group effect was found in the post hoc tests for Task 1, sentence recognition in quiet, $F(2, 30) = 12.106$, $p < .001$, $\eta^2 = .447$, the same task that showed differential effects of learning relative to pretest. All other corrected effects were nonsignificant (*p*s > .007). The largest effect size estimate obtained for these nonsignificant comparisons was $\eta^2 = .165$, in Task 4 (audiovisual word recognition), which reflects a small influence of group on AV recognition accuracy that was not statistically significant.

In the only task demonstrating group differences, (therefore, Task 1) Tukey's honestly significant difference post hoc comparisons were performed to investigate pairwise group differences. In these tests, recognition accuracy was found to be significantly better for the interactive (75.0%) and lab (81.1%) groups than for the control (59.3%; both *p* values < .005), whereas the interactive and lab groups did not differ significantly ($p = .389$).

One other measure was calculated for the results of the posttest tasks: AV benefit scores. AV benefit expresses the difference between each subject's A-only and AV recognition scores as a function of their initial audio-only accuracy: (AV accuracy) – (A-only accuracy)/1 – (A-only accuracy; Sumby & Pollack, 1954). That is, if a subject recognized audio-only words with 80% accuracy and gained 10% in the audiovisual condition, the gain would be greater relative to the potential for improvement than for someone who showed

**Figure 4.** Boxplots of the data from all six posttest tasks in Experiment 1, split by training group (see Figure 2 for details on boxplot parameters). The following tasks were completed in Experiment 1: Task 1 (sentence recognition in quiet, 70 dB presentation level), Task 2 (sentence recognition in noise, +8 dB signal-to-noise ratio [SNR]), Task 3 (audio-only word recognition), Task 4 (audio-visual [AV] word recognition), Task 5 (prosodic contour discrimination), and Task 6 (environmental sound recognition). Groups differed significantly on Task 1, but no other task showed a significant main effect of group on performance. Pairwise group contrasts in Task 1 were significant for the interactive/control and lab/control comparisons.



a 10% gain from 20% A-only to 30% in AV recognition. This measure therefore reflects AV integration gains more accurately than either A-only or AV scores alone. When AV benefit scores calculated for the present experiment were analyzed via corrected one-way ANOVA, however, group effects were also nonsignificant, $F(2, 30) = 2.939$, $p = .068$, $\eta^2 = .164$; α level of $p = .007$.

## *Discussion*

The first goal of Experiment 1, determining whether perceptual learning was possible through interactive exposure to vocoded acoustics, was satisfactorily met by these results. Subjects in the interactive training group showed significant improvements in their ability to recognize sentence-level speech in quiet, as predicted. Despite the differences between their training experience and the sentence-recognition-in-quiet test, this perceptual learning is perhaps not surprising, given other instances of generalization from novel speech or nonspeech training materials to a speech-in-quiet task (e.g., Loebach & Pisoni, 2008; Shafiro et al., 2012). Unlike previous studies showing training improvements on this task, however, the interactive training used here did not involve any explicit instruction or transcription. Subjects in the interactive training group did not receive exposure to such tasks, or any similar tasks, during their training period, yet their performance on this task improved relative to pretest. We interpret such improvement as an instance of *far transfer* of training effects (Ellis, 1965; Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, 2014), similar in kind to previously reported instances of transfer (e.g., from environmental sound training to sentence recognition [Shafiro et al., 2012]), but perhaps even greater in the dissimilarity between training and pretest–posttest assessment.

Although this improvement is encouraging, however, it constituted only one of our goals in designing the present experiment. The second goal was to determine whether the direct experience with perceptual skills such as audiovisual integration and speech perception in noise would result in broader perceptual gains for interactive subjects (relative to lab-trained subjects and controls). Our prediction that such gains would be observed was not met. No significant differences were observed across the three training groups in any of our additional posttest tasks: Subjects' perceptual skills in every task other than sentence recognition in the quiet were the same whether they had spent time in vocoded conversation, transcribing vocoded sentences, or not hearing vocoded acoustics at all.

These null results across five out of six posttest tasks were unexpected. In earlier studies, standard lab-based auditory training has been shown to generalize from training on sentence-level materials to testing on isolated words (Loebach & Pisoni, 2008), which our lab subjects did not (i.e., their performance on Task 3, isolated A-only word recognition, was not distinct from interactive or control subjects' performance). Furthermore, our interactive subjects had multiple opportunities during their training to engage in and improve their perceptual skills in the remaining domains: speech perception in noise (Task 2), AV integration (Task 4), recognition of speech prosody (Task 5), and recognition of environmental sounds (Task 6). Subjects in the lab and control groups had no experience with these aspects of vocoded perception, yet their performance following training was indistinguishable with that of interactive subjects.

There are three possible interpretations of these null results. First, it could be that our training methods genuinely do not result in learning for any domain other than sentence recognition in quiet. If this is so, then the next logical goal would be to understand why the experience of interactive subjects in domains such as audiovisual recognition does not result in perceptual benefits. A second alternative is that interactive subjects may have gained some differential perceptual benefits from their training, but our posttest assessments were not sufficient to reflect these gains. Experiment 2 constitutes an exploration of this possibility. A third possible interpretation lies between the two previous accounts: Interactive training may have the potential to result in generalized perceptual gains but for some reason did not have this effect in the present experiment. Possible explanations for this lack of effect might include the relatively short 1-hr training duration (cf. Shafiro et al., 2012; Smalt, Gonzalez-Castillo, Talavage, Pisoni, & Svirsky, 2013), the complex learning environment present in interactive training, or a combination of the two. This third intermediate interpretation will be explored in more depth in the General Discussion.

## Experiment 2

A second preliminary experiment was designed to test the possibility that the null result obtained in Experiment 1 for tasks other than sentence recognition in quiet was because of the nature of the particular set of posttest assessments, rather than a lack of perceptual gains on the part of interactive subjects relative to lab-trained subjects and controls. Experiment 2 therefore follows the same training protocol and involves the same basic design, looking for group differences in a variety of posttest tasks, but the particulars of these tasks were changed. Two types of changes were made: First, the sentence recognition in quiet and sentence recognition in noise tasks were modified from their Experiment 1 versions. Second, three new tasks, testing perceptual skills not measured in Experiment 1, were added.

In Experiment 1, there was an inequality of variance observed in the posttest results for the sentence recognition in quiet task (Task 1): Lab subjects were more consistent relative to interactive and control subjects (Figure 4). This difference may have been related to a ceiling effect, a limit on the accuracy with which subjects could recognize these vocoded IEEE sentences. Indeed, mean posttest sentence recognition scores of 75.0% and 81.1% are quite high, relative to reports in the literature (e.g., Loebach & Pisoni, 2008; Loebach et al., 2010; but see Davis et al., 2005). In Experiment 2, therefore, the difficulty of this task was increased by reducing the signal intensity of the stimuli by 5 dB to determine whether the difference in response variability was because of a ceiling effect or a genuine group difference. Believing the former to be more likely, we predicted that this change would eliminate the response variability difference between groups on this task.

The stimuli used in Task 2 (sentence recognition in noise) were altered to make recognition easier, improving average performance. Although there was no inequality of variance observed in this task in Experiment 1, overall performance was very poor, with 10 of 33 subjects recognizing fewer than 10% of the key words in these stimuli. Such low levels of performance potentially limit our ability to observe fine-grained differences in subjects' skills; a range of abilities could result in similar near-total failure to perform the task. The SNR of the stimuli in this task was therefore improved, from +8 dB in Experiment 1 to +10 dB in the current experiment's Task 2, and also to +15 dB SNR in the current experiment's Task 4. These tasks therefore provided two additional opportunities to test our prediction that group differences would be observed in speech-in-noise perception if subjects' average scores were further from a floor level of performance.

Three additional perceptual tasks were also added to Experiment 2. The first was a high-variability sentence recognition task (Task 3), where every sentence was produced by a different talker. Success at such a task requires more robust perceptual representations and adaptive recognition skills than at a single-talker task (Gilbert, Tamati, & Pisoni, 2013; Tamati, Gilbert, & Pisoni, 2013). If the challenges of interactive training resulted in improvements to indexical perception, subjects in the interactive group would do better on this high-variability task than their lab or control counterparts. The remaining two additional tasks departed from previous measure more substantially: Rather than being computer-based, these tests were both *live-voice* word recognition tasks, where subjects transcribed words spoken in real-time by the experimenter. In one test (Task 5), the target words were preceded by a standard carrier phrase, whereas in the other (Task 6), target words were preceded by a semantically related *context* phrase (see Method). Direct comparison between the two tasks, therefore, allowed for a measure of *context benefit* analogous to AV benefit analyzed in Experiment 1. Because interactive training potentially allowed subjects to use semantic context to aid in their word recognition to a greater degree than the standard lab training, we predicted that interactive subjects would show greater context benefit than subjects in either of the other two groups.

We chose to conduct these tasks live voice, rather than using prerecorded stimuli to include posttest assessments that were more similar to the training experiences of the interactive group than those of the standard lab group. The substantial differences between the training and test circumstances for interactive subjects mean that any success they have on these computer-based posttests would necessarily be examples of transfer of their training (cf. Tidwell et al., 2014). We believe that tightly controlled circumstances are necessary for strict, replicable tests of perceptual ability, and recognize that our live-voice tasks lose a critical measure of this control by having the experimenter produce the stimuli while aware of the training conditions experienced by each subject. Yet these tasks strike a compromise between the relatively free-form conditions of the interactive training and the consistency and control needed to conduct a quantitative assessment of perception. They therefore provide a strong test of the learning (potentially) achieved by interactive subjects: If interactive-training subjects cannot outperform lab and control subjects on these tasks (which would require transfer of lab and control computer-based training), then it is likely that further exploration of experience-related tasks (AV recognition, environmental sound recognition, etc.) would be unsuccessful as well.

## Method

All of the methods were approved for use with human subjects by the Indiana University Institutional Review Board.

*Subjects.* Twenty-one volunteers, nine men and 12 women, were recruited from the Indiana University community and paid for their participation. No volunteers who were tested in Experiment 1 were included in Experiment 2. Subjects were all monolingual speakers of North American English between the ages of 18 and 35 years who reported no history of speech or hearing problems and passed an audiological screener with thresholds at or below 25 dB SPL (NH, < 25 dB HL).

*Tasks and stimuli.* The following tasks were included in Experiment 2, in the following order: sentence recognition in quiet (Task 1), sentence recognition in noise at +10 dB SNR (Task 2), high-variability sentence recognition (Task 3), sentence recognition in noise at +15 dB SNR (Task 4), live-voice isolated word recognition (Task 5), and live-voice word recognition in semantic context (Task 6). Tasks 1 and 2 were completed as both pretest and posttest measures (see Figure 5). Training and between-groups design were the same as in Experiment 1.

Task 1 (sentence recognition in quiet) presented subjects with 20 unique IEEE sentences in random order. Each sentence contained five keywords, for a total of 100 per iteration of the task. Stimuli were produced by a single female talker from the Midwest dialect region and were presented at an average of 65 dB SPL – 5 dB lower than the presentation level of stimuli from the comparable task in Experiment 1.

In the pre- and posttest iterations of Task 2, subjects were asked to transcribe 20 unique HINT sentences in the presence of competing multitalker babble. Targets were presented at a +10 dB SNR, a 3-dB difference from Experiment 1, and were combined with the four-talker babble track using the methods described. Sentence stimuli contained between two and five keywords each for a total of 73 in each version of the task. Task 4 was identical to Task 2 except that stimuli were a novel set of 20 randomly selected HINT sentences, and they were presented at +15 dB SNR.

Task 3 (high-variability sentence recognition) once again used 20 IEEE stimuli. Each stimulus was produced by a different randomly selected talker from the Hoosier Multitalker Sentence Database (Karl & Pisoni, 1994). Stimuli from these 20 talkers, 10 men and 10 women, were digitally leveled to have the same average SPL and were presented in random order at 70 dB SPL.

During Tasks 5 and 6 (live-voice isolated word recognition), the experimenter (a female native speaker of Midwestern North American English) presented subjects with spoken words from the Central Institute of the Deaf W-22 audiological test set (Hirsh, Davis, Silverman, Reynolds, Eldert, & Benson, 1952)—List 3A for Task 5 and List 4A for Task 6. These items were all monosyllabic words with high frequency and comprehensive inclusion of the phonetic segments of English. Target words in each trial were preceded by a short, carrier phrase. In Task 5, this phrase was always "*The next target word is...*" In Task 6, each target word was paired with a short semantically related sentence. These *context clue* sentences were generated for the purposes of this experiment, and achieved relatedness in a variety of ways, including mention of members of the same semantic category (e.g., "*August and September are booked up*" with target *May*), highly associated concepts (e.g., "*Puppies sleep in piles*" with target *cute*), synonyms (e.g., "*Boring plays feel long*" with target *dull*), and antonyms ("*Please tell me the truth*" with target *lie*). All context sentences (Tasks 5 & 6) were between four and six words in length and were spoken with a short pause (approximately 1 s) intervening between the end of the phrase and the target word. The full list of context sentences and target words are available in the online supplemental materials (see Supplemental Table S1). Subjects in all three training groups responded to the same items in Tasks 5 and 6, and the critical comparison for these tasks was the difference in recognition performance achieved between them; therefore, the materials were not normed with respect to effective cueing, as only between-subjects differences among the items would be analyzed (see Data Analysis).

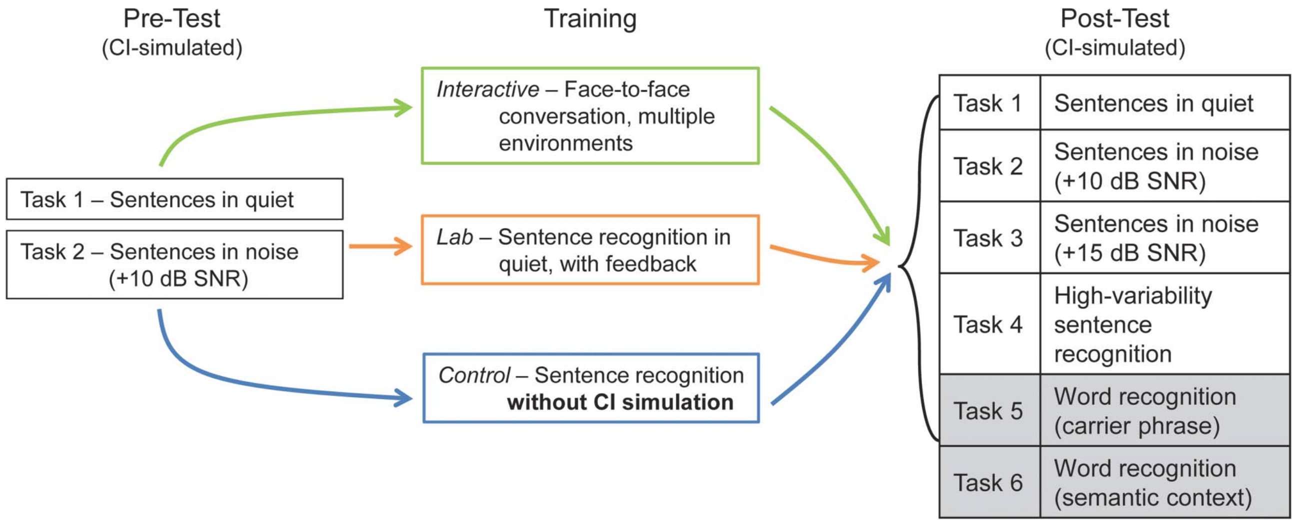
## Procedure

The procedure for this experiment was identical to the methods used in Experiment 1. Subjects were randomly assigned to one of the three exposure groups and given a pure-tone audiological screener. They were then fitted with the PRTV and seated in the same setting used in Experiment 1: A sound-attenuating booth in front of a computer monitor and keyboard. All subjects then completed the two pretest versions of Tasks 1 and 2 and were given 1 hr of experience with the PRTV acoustic transformation. The nature of these training periods varied across groups, with the same procedure, materials, and design as described for Experiment 1. Following 1 hr of training, each subject completed the battery of six posttest tasks. The first four posttests were conducted in the sound-attenuating booth using the computer interface; Tasks 5 and 6 were conducted outside the booth in the adjacent testing room, previously used for the informed consent process. For these two live-voice tasks, the experimenter read aloud a list of isolated English words, and the subjects transcribed them, writing their responses on a pen-and-paper answer sheet. Both parties were seated approximately six feet apart, and no obstruction was placed between the experimenter and the subject. Visual information was therefore fully available during both word recognition tasks.

## Data Analysis

Perceptual accuracy was calculated for each pre- and posttest task. Sentence transcription tasks were scored via

**Figure 5.** Schematic of the design of Experiment 2. Training protocols were identical to Experiment 1, but pre- and posttest tasks were either modified from Experiment 1 (Tasks 1, 2, & 4) or assessed a novel perceptual skill (Tasks 3, 5, & 6). Shaded boxes indicate live-voice testing modality (Tasks 5 & 6). All subjects completed identical pre- and posttest tasks.



Pre-Test
(CI-simulated)
Task 1 – Sentences in quiet
Task 2 – Sentences in noise (+10 dB SNR)
Training
Interactive – Face-to-face conversation, multiple environments
Lab – Sentence recognition in quiet, with feedback
Control – Sentence recognition without CI simulation
Post-Test
(CI-simulated)
Task 1 Sentences in quiet
Task 2 Sentences in noise (+10 dB SNR)
Task 3 Sentences in noise (+15 dB SNR)
Task 4 High-variability sentence recognition
Task 5 Word recognition (carrier phrase)
Task 6 Word recognition (semantic context)

the number of keywords correctly recognized, with half credit again awarded for responses deviating from a target by a single segment or by number/tense morphology. Accuracy in the word recognition tasks was assessed with binary scoring, where each response was either correct in its entirety or it received no credit. To assess the effects of context on word recognition, a measure of context benefit analogous to AV benefit was also calculated. Because of the relatively small sample size of this follow-up study and in the interest of comparison with Experiment 1 and future work, both standard statistical results and effect size analyses will be reported in the Results section.

### Results

*Initial equivalency.* Separate one-way ANOVAs, corrected for multiple comparisons across these and the analyses in the Comparison Across Exposure Types section, were run on the recognition-in-quiet and recognition-in-babble pretests to determine initial equivalency among subjects in the three training groups. The effect of group was not significant in either task (*p*s > .0055, corrected α), and the estimated effect sizes were also very small (Task 1, $\eta_p^2$ = .041; Task 2, $\eta_p^2$ = .064). There were no significant differences, therefore, among groups prior to training.

*Comparison across exposure types.* As in Experiment 1, one-way corrected ANOVAs were conducted on each of the posttest perceptual tasks (Figure 6). The results of these analyses are summarized in Table 1. Correction for nine comparisons (including tests of initial group equivalency and all seven posttest task analyses) was applied to all tests in determining significance (Bonferroni correction, α = .0055). The homoscedasticity assumption was met in for the between-groups data for all analyses except Task 6, live-voice word recognition in semantic context (Levene's test, $p$ = .010), where there was significantly less variability in the interactive subject scores than those of the other two groups.

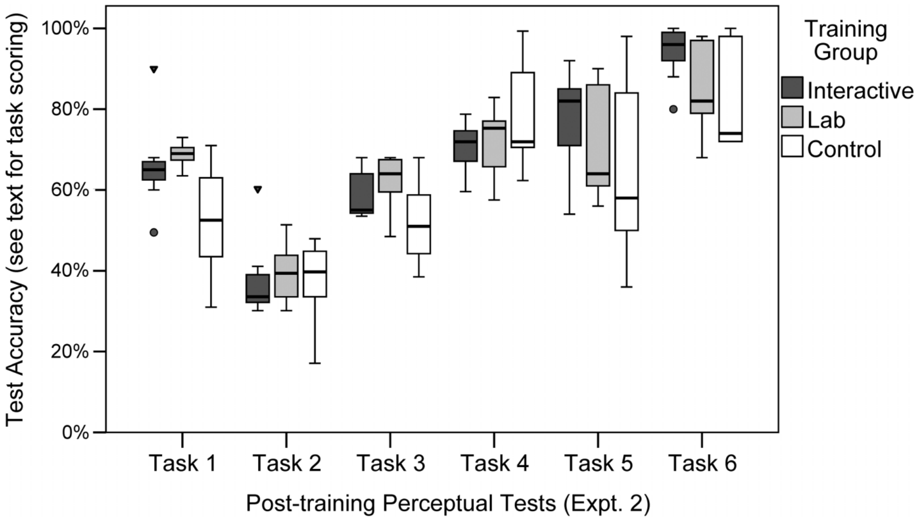
As the Table 1 summary shows, none of the ANOVA returned significant effects of group on perceptual accuracy. Effect size estimates for these nonsignificant tests provide some additional information regarding the potential for further exploration but are by no means substitutes for reliable effects in a statistical test. A moderate effect size of $\eta_p^2$ = .327 was observed in Task 1 (sentence recognition in quiet), and a small effect size of $\eta_p^2$ = .223 was estimated for Task 3 (high-variability sentence recognition). These two tasks did not show significant effects of training group; effect sizes are reported for completeness, especially given the small sample size.

As in Experiment 1, an additional analysis of context benefit was performed, examining scores on Tasks 5 and 6 in more detail. Corrected ANOVA on these context benefit data (Figure 7) did not show a significant effect of group, $F(2, 19) = 2.086$, $p = .153$, $\eta_p^2 = .188$.

Given the importance of the comparisons between interactive and lab subjects and the motivation for live-voice testing described, preplanned comparisons were also conducted between the interactive and lab groups on these tasks (Tasks 5 and 6) and their derived context benefit scores. These planned independent samples $t$ tests were not significant after correction for multiple comparisons, Task 5: $t(12) = 0.695$, $p = .501$, Hedge's $g = 0.347$; Task 6: $t(12) = 1.580$, $p = .140$, Hedge's $g = 0.792$; context benefit: $t(12) = 2.269$, $p = .043$, Hedge's $g = 1.13$.

It should be noted, however, that despite the lack of corrected statistical significance, the effect size estimates for these interactive/lab comparisons range from moderate, in the case of Task 5, to very large, for context benefit (Hedge's $g$ is an effect size measure analogous to Cohen's $d$ but which corrects for the bias observed in Cohen's $d$ for small sample

**Figure 6.** Boxplots of the data from all six posttest tasks in Experiment 2, split by training group (see Figure 2 for details on boxplot parameters). In Experiment 2, the following tasks were completed: Task 1 (sentence recognition in quiet, 65 dB presentation level), Task 2 (sentence recognition in noise, +10 dB signal-to-noise ratio [SNR]), Task 3 (high-variability sentence recognition), Task 4 (sentence recognition in noise, +15 dB SNR), Task 5 (live-voice word recognition in a carrier phrase), and Task 6 (live-voice word recognition with semantic context). No significant effects of group were found for these tasks.



sizes; Lakens, 2013). In the case of context benefit scores, for example, a Hedge's *g* of 1.13 corresponds to an 80% chance that random samples from similar populations would replicate the observed difference in context benefit (common language effect size, cf. Lakens, 2013). Although these differences were not statistically significant, we report them here to document the large effect size of the planned interactive/lab training comparison. However, it is also important to recall that neither the interactive nor lab groups were significantly different from the control in terms of context benefit or word recognition accuracy on Tasks 5 and 6. Such contrast with the performance of the no-training control group is critical for any result to be interpreted as a straightforward effect of training, which we cannot do here.

## Discussion

The lack of significant training group differences on the posttest tasks in Experiment 2 suggests that the design and/or selection of particular tests was not responsible for the null results obtained here or in Experiment 1. The design goals for the sentence recognition in quiet task (Task 1) and sentence recognition in noise (Tasks 2 and 4) were achieved: Accuracy variance did not differ across groups in this version of Task 1, and inspection of Figure 6 confirms that performance on Tasks 2 and 4 was well above the range where floor effects would potentially reduce power. These changes did not, however, result in significant differences in performance between the interactive subjects and those in the lab and control groups, as predicted.

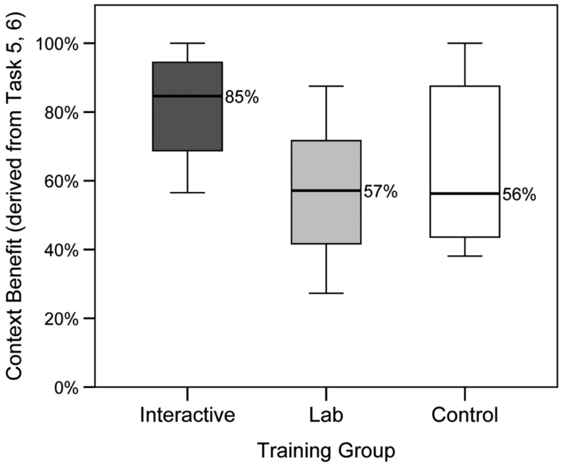
The lack of differences observed in the tasks assessing new perceptual domains (perception of high-variability stimuli and live-voice recognition) provide further support that the particulars of individual tests are not responsible for this pattern of results. These tasks assessed skills related specifically to the training experiences of interactive subjects, yet these subjects showed no benefits relative to lab subjects or controls. This was particularly surprising in the case of Tasks 5 and 6, in which interactive subjects were responding to a talker whose vocoded voice was highly familiar

**Table 1.** Summary of posttest analysis of variance for main effects of group from Experiment 2.

| Posttest measure | *F* | *p* | $\eta_p^2$ |
|---|---|---|---|
| Task 1 (sentence recognition in quiet, 65 dB presentation level) | 4.381 | .028 | .327 |
| Task 2 (sentence recognition in noise, +10 dB SNR) | 0.075 | .928 | .008 |
| Task 3 (high-variability sentence recognition) | 2.583 | .103 | .223 |
| Task 4 (sentence recognition in noise, +15 dB SNR) | 1.680 | .214 | .157 |
| Task 5 (live-voice isolated word recognition) | 0.753 | .485 | .077 |
| Task 6 (live-voice word recognition with context) | 1.600 | .229 | .151 |
| Context benefit score (derived from Tasks 5 & 6) | 2.086 | .153 | .188 |

*Note.* *P* values should be interpreted relative to α = .0055 (Bonferroni correction for multiple comparisons). SNR = signal-to-noise ratio.

**Figure 7.** Boxplots of the derived context benefit scores for each training group (see Figure 2 for boxplot parameter details). Context benefit expresses the improvement observed in each subject's word recognition from Task 5 (no context) to Task 6 (with context) as a percentage of the total possible improvement. No significant group effects were observed. Planned comparisons between the interactive and lab groups were also nonsignificant, although effect size estimates for this contrast were large (see Results).



and in the same context in which they were trained (face-to-face interaction), but lab and control subjects were responding to a novel vocoded voice in an unfamiliar context. In light of these differing training experiences, the finding that interactive subjects show gains relative to controls in a computer-based sentence recognition in quiet task with an unfamiliar talker (Experiment 1), but not on a live-voice recognition task with a familiar talker, is puzzling. Achieving transfer of training to performance on untrained tasks has a long history of recognition as a challenging goal (Ferguson et al., 2014; Loebach et al., 2009; Tidwell et al., 2014), and the results of Experiment 2 suggest that producing such robust learning in the domain of speech is not an exception to this difficulty.

## General Discussion

Following Experiment 1, we suggested three possible interpretations for the lack of differences observed among the interactive, lab, and control training groups on the majority of posttest tasks: First, that interactive training does not result in improvements in domains other than sentence recognition in quiet; second, that improvements can and did occur in other domains, but the particular tasks selected did not access or reflect these skills; or third, that interactive training could result in improvements beyond sentence recognition in quiet but that there were factors involved in the training that limited these improvements. The results of Experiment 2 greatly reduced the likelihood of the second interpretation, but the remaining two possibilities bear further discussion.

For reasons stated in the motivation and discussion of Experiments 1 and 2, we believe it is highly unlikely that interactive training of the kind used here would result in improvements restricted only to the domain of sentence recognition in quiet. Succinctly put, the interactive training involves much more than the skills needed to succeed at this one task. There would have to be powerful learning biases at play—biases we see no evidence of in similar work (e.g., Loebach & Pisoni, 2008; Shafiro et al., 2012)—for such training to benefit only the limited set of speech-in-quiet skills and no other domains.

Moreover, there are excellent additional reasons to believe that interactive training should promote improved perceptual learning in these other domains. Interactive training allows subjects to hear their own speech through the vocoding transformation in real time, to test hypotheses about the acoustic world to which they are being exposed, and to actively explore the range of acoustic possibilities for themselves and their immediate environments. A rich literature in perceptual and motor learning has demonstrated that such active engagement in exploring the perceptual world is critical for achieving learning that is both accurate and robust (e.g., Bingham, 1988; Feldman, 1966; Haidet, Morgan, O'Malley, Moran, & Richards, 2004; Held & Hein, 1963). Active engagement on the part of subjects has been shown to be critical in the learning of motor skills, and the greater the degree of participation, the more generalizable any gains become (e.g., Feldman, 1966; Snapp-Childs, Casserly, Mon-Williams, & Bingham, 2013). These domain-general findings have affected the development of applied training strategies nearly across the board, in situations as varied as physical therapy clinics, middle school classrooms, and athletic instruction (Bonwell & Eison, 1991; Hogan et al., 2006; Kramer & Erickson, 2007; Kwakkel, Kollen, & Krebs, 2008; Williams & Hodges, 2005). Although this study failed to find any similar global benefits for interactive training on CI-simulated auditory perception, such benefits are likely to exist in some domain or following different interactive training experiences or activities.

If the potential for such interactive training therefore remains high, the only explanation left to us for the lack of transfer in the present experiments is that some factor or factors must have limited the potential for learning during training. One possible candidate is the relatively short duration of the training protocol used here. Although it is comparable to the length of training in other standard lab-training studies (Davis et al., 2005; Loebach & Pisoni, 2008; Loebach et al., 2010), the only other study that has examined the effects of fully interactive, freely moving exposure to CI-simulated signals had subjects complete 2 hr of training every day for 2 weeks (Smalt et al., 2013)—an order of magnitude beyond the 1 hr subjects spent in training in the present work. Future studies involving multiple experimental sessions, potentially with longer periods of training within each session, could test the impact of this variable on performance.

Another possible factor limiting the learning gained by subjects in the present study concerns the high complexity and variability of the interactive training protocol. To give subjects a breadth of experience and maximize the differences between interactive and standard lab training, subjects in the interactive condition were given exposure to no fewer than four different environmental conditions (e.g., reverberant hallway, student lounge, sidewalk outdoors, etc.), diverse environmental sounds, and multiple sources of speech and nonspeech background noise. Although this high degree of variability was desirable in some respects, it may have spread the potential for observing robust learning too thin, applying only minimally to any one domain or skill.

There is a large literature on what has been called *high-variability phonetic training* (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow & Bent, 2008; Clopper & Pisoni, 2004; Logan, Lively, & Pisoni, 1991; Sidaras, Alexander, & Nygaard, 2009). This method of linguistic training produces generalized gains in skills such as recognition of nonnative phonemic contrasts (Bradlow et al., 1999; Logan et al., 1991), understanding accented speech (Bradlow & Bent 2008; Clopper & Pisoni, 2004; Sidaras et al., 2009), and accuracy in speech articulation (Bradlow et al., 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). The central tenet of high-variability phonetic training is that, in order to create nuanced, generalizable learning, subjects must be trained on a wide variety of materials: Different talkers, different degrees and types of foreign accent, and different semantic and phonological contexts, and so forth. The learning that results from such high-variability training is able to be generalized to novel materials and situations—but critically for the present discussion, achieving that robust learning is initially slower and more difficult for subjects than in conventional low-variability training protocols. For example, a subject could learn the speaker-specific characteristics of Mandarin-accented English speech in a single short session, but would need multiple sessions with a variety of talkers and lexical items to gain robust, talker-general, and context-independent representations of this nonnative variety of English (Bradlow & Bent, 2008).

By extension, subjects in our interactive training may have been hampered in their learning efficiency by the degree of variability in their training experience. If the comparison to high-variability (laboratory-based) phonetic training is accurate, then our subjects would have eventually achieved not only observable perceptual gains but also particularly robust and generalizable gains. In addition, this explanation predicts that learning could be obtained from a protocol similar to that of Experiments 1 and 2 but modified in one of two ways: either by allowing subjects to engage in longer periods of training or by reducing some aspects of the variability present in the training (e.g., eliminating the speech-in-noise or multiple acoustic environments aspects). We tentatively suggest that the later strategy may be more fruitful, because real-world CI users (who have been very thoroughly trained in these situations through standard usage) typically continue to show large deficits to speech perception in noisy or challenging environments (e.g., Friesen, Shannon, Baskent, & Wang, 2001). Such predictions provide concrete targets for future work exploring interactive auditory training of this nature.

## Conclusions

Overall, the present set of experiments constitutes an initial step in our exploration of interactive real-time training with CI-simulated acoustic signals. Perceptual learning following interactive training was observed in sentence recognition in the quiet (Experiment 1), but no gains relative to lab subjects or controls were seen in other domains, even those in which interactive subjects received direct experience (e.g., perception of speech in noise, audiovisual word recognition). This lack of differential perceptual gains from interactive training appears not to be because of design or selection issues of the particular posttest measures used here (Experiment 2) but rather to reflect some inherent limitation on the benefits of this kind of auditory training in its current instantiation. Longer training periods or more narrowly restricted activities during training may alleviate these limitations, if comparisons to high-variability phonetic training are accurate. Although there appears to be potential for interactive training methods of this type to produce limited perceptual learning and transfer from interactive to computer-based tasks, a great deal concerning the effects of training and its effective transfer to untrained tasks clearly remains to be understood.

## Acknowledgments

This study was supported in part by the National Institute on Deafness and Other Communication Disorders, Grants R01-00111 and T32-DC00012 to David Pisoni at Indiana University. This research appeared as part of the doctoral dissertation of the first author. We thank Kate Sherwood for her dedication in data collection for Experiment 1, Geoffrey Bingham for advice during the transition between Experiments 1 and 2, and Charles Brandt for technical support.

## References

**Baskent, D., & Shannon, R. V.** (2003). Speech recognition under conditions of frequency-place compression and expansion. *The Journal of the Acoustical Society of America, 113*(4), 2064–2076.

**Bent, T., Buchwald, A., & Pisoni, D. B.** (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America, 126*(5), 2660–2669.

**Bingham, G.** (1988). Task-specific devices and the perceptual bottleneck. *Human Movement Science, 7,* 225–264.

**Bonwell, C. C., & Eison, J. A.** (1991). *Active learning: Creating excitement in the classroom.* Washington, DC: School of Education and Human Development, George Washington University.

**Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y.** (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985.

**Bradlow, A. R., & Bent, T.** (2008). Perceptual adaptation to nonnative speech. *Cognition, 106*(2), 707–729.

**Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y.** (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101* (4), 2299–2310.

**Carroll, J., & Zeng, F.** (2007). Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations. *Hearing Research, 231*(1–2), 42–53.

**Casserly, E.** (2013). *Effects of real-time cochlear implant simulation on speech perception and production* (unpublished doctoral thesis). Bloomington, IN: Indiana University.

**Casserly, E., Pisoni, D. B., Smalt, C. J., & Talavage, T.** (2011, May). *A portable, real-time vocoder: Technology and preliminary perceptual learning findings*. Paper presented at the 161st Meeting of the Acoustical Society of America, Seattle, WA.

**Clopper, C. G., & Pisoni, D. B.** (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47*(2), 207–238.

**Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C.** (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241.

**Dorman, M. F., Dankowski, K., McCandless, G., Parkin, J. L., & Smith, L. B.** (1991). Vowel and consonant recognition with the aid of a multichannel cochlear implant. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43*(A), 585–601.

**Ellis, H.** (1965). *The transfer of learning*. New York, NY: Macmillan.

**Feldman, A.** (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II. Controllable parameters of the muscles. *Biophysics, 11,* 565–578.

**Ferguson, M. A., Henshaw, H., Clark, D. P. A., & Moore, D. R.** (2014). Benefits of phoneme discrimination training in a randomized controlled trial of 50- to 74-year-olds with mild hearing loss. *Ear and Hearing, 35*(4), e110–e121.

**Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X.** (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America, 110*(2), 1150–1163.

**Fu, Q.-J., Nogaki, G., & Gavin, J. J.** (2005). Auditory training with spectrally shifted speech: Implications for cochlear implant patient auditory rehabilitation. *Journal of the Association for Research in Otolaryngology, 6*(1), 180–189.

**Fu, Q.-J., Shannon, R. V., & Wang, X.** (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America, 14*(6), 3586–3596.

**Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N., & Zue, V.** (1993). *The DARPA TIMIT acoustic-phonetic continuous speech corpus*. Philadelphia, PA: Linguistic Data Consortium.

**Geers, A., Brenner, C., & Davidson, L.** (2003). Factors associated with development of speech perception skills in children implanted by age five. *Ear and Hearing, 22*(1 Suppl.), 24S–35S.

**Geers, A., Tobey, E. A., & Moog, J. S.** (2011). Long-term outcomes of cochlear implantation in early childhood. *Ear and Hearing, 32*(1), 1S–92S.

**Gilbert, J. L., Tamati, T. N., & Pisoni, D. B.** (2013). Development, reliability and validity of PRESTO: A new high-variability sentence recognition test. *Journal of the American Academy of Audiology, 24*(1), 26–36.

**Glass, G. V., Peckham, P. D., & Sanders, J. R.** (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288.

**Haidet, P., Morgan, R. O., O'Malley, K., Moran, B. J., & Richards, B. F.** (2004). A controlled trial of active versus passive learning strategies in a large group setting. *Advances in Health Sciences Education, 9*(1), 15–27. doi:10.1023/b:ahse.0000012213.62043.45

**Held, R., & Hein, A.** (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology, 56*(5), 872–876.

**Henshaw, H., & Ferguson, M. A.** (2013). Efficacy of individual computer-based auditory training for people with hearing loss: A systematic review of the evidence. *PLoS One, 8,* e62836.

**Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W.** (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 15,* 321–337.

**Hogan, N., Krebs, H. I., Rohrer, B., Palazzolo, J. J., Dipietro, L., Fasoli, S. E., . . . Volpe, B. T.** (2006). Motion of muscles? Some behavioral factors underlying robotic assistance of motor recovery. *Journal of Rehabilitation Research & Development, 43*(5), 605–618.

**Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B.** (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46*(2), 390–404.

**Kaiser, A. R., & Svirsky, M.** (1999). A real time PC based cochlear implant speech processor with an interface to the nucleus 22 electrode cochlear implant and a filtered noiseband simulation. *Research on Spoken Language Processing: Progress Report No. 23* (pp. 417–428). Bloomington, IN: Indiana University.

**Kaiser, A. R., & Svirsky, M.** (2000). *Using a personal computer to perform real-time signal processing in cochlear implant research*. Paper presented at the IXth IEEE-DSP Workshop, Hunt, TX.

**Karl, J. R., & Pisoni, D. B.** (1994). The role of talker-specific information in memory for spoken sentences. *The Journal of the Acoustical Society of America, 95*(5), 2873.

**Kirk, K. I., Pisoni, D. B., & Osberger, M. J.** (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing, 16*(5),470–481.

**Kramer, A. F., & Erickson, K. I.** (2007). Capitalizing on cortical plasticity: Influence of physical activity on cognition and brain function. *Trends in Cognitive Sciences, 11*(8), 342–348.

**Kwakkel, G., Kollen, B. J., & Krebs, H. I.** (2008). Effects of robot-assisted therapy on upper limb recovery after stroke: A systematic review. *Neurorehabilitation and Neural Repair, 22*(2), 111–121.

**Lakens, D.** (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. doi:10.3389/fpsyg.2013.00863

**Loebach, J. L., Bent, T., & Pisoni, D. B.** (2008). Multiple routes to the perceptual learning of speech. *The Journal of the Acoustical Society of America, 124*(1), 552–561.

**Loebach, J. L., & Pisoni, D. B.** (2008). Perceptual learning of spectrally degraded speech and environmental sounds. *The Journal of the Acoustical Society of America, 123*(2), 1126–1139.

**Loebach, J. L., Pisoni, D. B., & Svirksy, M.** (2009). Transfer of auditory perceptual learning with spectrally reduced speech to speech and nonspeech tasks: Implications for cochlear implants. *Ear and Hearing, 30*(6), 662–674.

**Loebach, J. L., Pisoni, D. B., & Svirsky, M.** (2010). Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *Journal of Experimental Psychology: Human Perception & Performance, 36*(1), 224–234.

**Logan, J. S., Lively, S. E., & Pisoni, D. B.** (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874–886.

**Luce, P. A., & Pisoni, D. B.** (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing, 19*(1), 1–36.

**Nilsson, M., Soli, S. D., & Sullivan, J. A.** (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds. *The Journal of the Acoustical Society of America, 95*(2), 1085–1099.

**Nusbaum, H. C., Pisoni, D. B., & Davis, C. K.** (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

**Pisoni, D. B., Svirsky, M., Kirk, K. I., & Miyamoto, R. T.** (1997, May). *Looking at the "stars:" A first report on the intercorrelations among measures of speech perception, intelligibility, and language in pediatric cochlear implant users*. Paper presented at the 5th International Cochlear Implant Conference, New York, NY.

**Shafiro, V.** (2008a). Development of a large-item environmental sound test and the effects of short-term training with spectrally-degraded stimuli. *Ear and Hearing, 29*(5), 775–790. doi:10.1097/AUD.0b013e31817e08ea

**Shafiro, V.** (2008b). Identification of environmental sounds with varying spectral resolution. *Ear and Hearing, 29*(3), 401–420. doi:10.1097/AUD.0b013e31816a0cf1

**Shafiro, V., Sheft, S., Gygi, B., & Ho, K. T. N.** (2012). The influence of environmental sound training on the perception of spectrally degraded speech and other sounds. *Trends in Amplification, 16*(2), 83–101.

**Shannon, R. V., Zeng, F., Kamath, V., & Ekelid, M.** (1995, October 13). Speech recognition with primary temporal cues. *Science, 270*(13), 303–304.

**Sheffert, S. M., Lachs, L., & Hernandez, L. R.** (1996). The Hoosier audiovisual multitalker database. *Research on Spoken Language Processing Progress Report, 21,* 578–583.

**Sidaras, S. K., Alexander, J. E. D., & Nygaard, L. C.** (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America, 125* (5), 3306–3316.

**Smalt, C. J., Gonzalez-Castillo, J., Talavage, T. M., Pisoni, D. B., & Svirksy, M. A.** (2013). Neural correlates of adaptation in freely-moving normal hearing subjects under cochlear implant acoustic simulations. *NeuroImage, 82,* 500–509.

**Snapp-Childs, W., Casserly, E., Mon-Williams, M., & Bingham, G.** (2013). Active prospective control is required for effective sensorimotor learning. *PLoS One, 8*(10), e77609. doi:10.1371/journal.pone.0077609

**Stelmachowicz, P. G., Hoover, B. M., Lewis, D. E., Kortekaas, R. W. L., & Pittman, A. L.** (2000). The relation between stimulus context, speech audibility, and perception for normal-hearing and hearing-impaired children. *Journal of Speech, Language, and Hearing Research, 43,* 902–914.

**Stenfelt, S., & Håkansson, B.** (2002). Air versus bone conduction: An equal loudness investigation. *Hearing Research, 167,* 1–12.

**Sumby, W. H., & Pollack, I.** (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215.

**Tamati, T. N., Gilbert, J. L., & Pisoni, D. B.** (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology, 24*(7), 616–634.

**Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L.** (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin and Review, 21,* 620–628.

**Williams, A. M., & Hodges, N. J.** (2005). Practice, instruction and skill acquisition in soccer: Challenging tradition. *Journal of Sports Sciences, 23*(6), 637–650. doi:10.1080/02640410400021328