# Individual Differences in the Perceptual Learning of Degraded Speech:

# Implications for Cochlear Implant Aural Rehabilitation

by

Augusta Katherine Helms Tillery

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2015 by the
Graduate Supervisory Committee:

Julie M. Liss, Chair
Tamiko Azuma
Christopher A. Brown
Michael F. Dorman
Rene L. Utianski

ARIZONA STATE UNIVERSITY

May 2015

# ABSTRACT

In the noise and commotion of daily life, people achieve effective communication partly because spoken messages are replete with redundant information. Listeners exploit available contextual, linguistic, phonemic, and prosodic cues to decipher degraded speech. When other cues are absent or ambiguous, phonemic and prosodic cues are particularly important because they help identify word boundaries, a process known as lexical segmentation. Individuals vary in the degree to which they rely on phonemic or prosodic cues for lexical segmentation in degraded conditions.

Deafened individuals who use a cochlear implant have diminished access to fine frequency information in the speech signal, and show resulting difficulty perceiving phonemic and prosodic cues. Auditory training on phonemic elements improves word recognition for some listeners. Little is known, however, about the potential benefits of prosodic training, or the degree to which individual differences in cue use affect outcomes.

The present study used simulated cochlear implant stimulation to examine the effects of phonemic and prosodic training on lexical segmentation. Participants completed targeted training with either phonemic or prosodic cues, and received passive exposure to the non-targeted cue. Results show that acuity to the targeted cue improved after training. In addition, both targeted attention and passive exposure to prosodic features led to increased use of these cues for lexical segmentation. Individual differences in degree and source of benefit point to the importance of personalizing clinical intervention to increase flexible use of a range of perceptual strategies for understanding speech.

ACKNOWLEDGMENTS

My long-term interest as a speech-language pathologist has been to help people with hearing loss achieve effective communication in their daily lives. After working for several years with cochlear implant recipients, I stepped into the realm of speech, language and hearing research, with the aim of strengthening my understanding of the perceptual and technological factors affecting communication for this group of individuals. My doctoral training has involved a circuitous but worthwhile journey through the disciplines of environmental acoustics, auditory physiology, psychophysics, speech science, signal processing, multisensory perception, and cognition. Of the many lessons I have learned, one of the most critical has been that scientific endeavors are strengthened by effective collaborations. My dissertation research has been no exception.

Without the support of my committee, colleagues, friends and family, this project would not have been realized. Julie Liss, my faculty advisor, welcomed me into her lab and supplied crucial support in the forms of conceptual guidance and research assistants. Her gift for visualizing the big picture served to keep my detail-driven ruminations in check. Rene Utianski, whose translational brilliance helped shape the big picture into concrete form, has been an unflinching source of positive encouragement. Michael Dorman, in whose Speech Science class I first encountered cochlear implants 27 years ago, is a model for conveying research ideas with clarity and concision. He also supplied critical resources for software purchasing and subject reimbursement. Tamiko Azuma's exacting advice on data analysis and manuscript preparation improved this document's organization and readability. Christopher Brown, whose generous mentoring from Day One has not ceased with time and distance, taught me important lessons in everything

from psychoacoustics and signal processing to the technicalities of pulling a really good espresso shot. I am also indebted to Sid Bacon, who originally accepted me into his lab, and always set aside time to explain the intricacies of the auditory periphery.

Many fellow students and colleagues have been instrumental in the development and completion of this project. Cimarron Ludwig, Stephanie Borrie, Sarah Cook, Christine Delfino, Elizabeth Fall, Saul Frankford, Taylor Hickok, Anbar Najam, Danielle Samuels, Lena Sarsour, Steven Sandoval, Hilda Torres, and others I have unintentionally omitted have cumulatively provided countless hours of assistance with experimental design, audio recording, stimulus generation, programming, subject recruitment, data collection, and transcription coding. I am also thankful for Bill Yost's willingness to share his space, resources, and prodigious knowledge of binaural hearing.

The support of family and friends was no less essential. Wednesday lunches, weekend dinners, Sunday teatimes, summer camping trips, and visits with the Aunties have provided regular respites from the cares of grad school, and are reminders of the things that really matter. My grandmothers, both gone now, were inspiration to persevere even when the going got tough. I am also beholden to our friend and neighbor Greg for lending his north Idaho lakeside haven for writing and reflection.

Above all, I am forever grateful to my husband Steve for his tremendous love and support. His enthusiastic and generous endorsement of my return to school made this journey possible. Beyond that, his ability to fill multiple roles as my better half, DIY home improvement specialist, killer string player, research guru, MATLAB programmer, and statistician is unparalleled. I look forward to many more shared adventures in life and science.

This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health, under award numbers R01DC008329, R01DC006859, and R21DC012558-01.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

# Introduction

In the everyday lives of most individuals, communication occurs despite such obstacles as background noise, garbled signals, and environmental distractions. Listeners accomplish this by leveraging the abundance of redundant information in connected speech: cues from the contextual, syntactic, semantic, lexical, phonetic, and prosodic domains are variably exploited in the process of deciphering a degraded utterance.

Many of these cues are broadly characterized as occupying either the segmental or the suprasegmental level of speech. At the segmental level, strings of phonemes form words, phrases, and sentences. Extending through these segmental strings are variations in fundamental frequency (F0), amplitude, and syllable duration. Collectively referred to as prosody, these suprasegmental variations serve multiple linguistic functions, many of which reinforce the information carried by segmental units.[1] For example, yes-no questions can be conveyed simultaneously by both word order and prosodic changes; information about lexical boundaries in connected speech is found both in phonemic units and in prosodic markers to syllabic stress; the emotional state of the speaker can be gleaned from the words uttered as well as tone of voice.

Of the functions served by segmental and suprasegmental cues, the one thought to be particularly important to the perception of degraded speech is the marking of word boundaries (Cutler and Butterfield, 1992; Liss et al., 1998; Mattys et al., 2005). Mattys et al. (2005) proposed a hierarchical model of speech segmentation, in which the effortless

[1] Some suprasegmental cues provide information that complements, rather than reinforces, the segmental cues. Irony, for example, is conveyed by manipulating prosody to convey a meaning that is opposite the one carried by the words.

parsing of continuous speech in optimal conditions gives way to progressively greater reliance on first segmental and then suprasegmental cues as the amount of contextual information diminishes and the degree of signal degradation increases. Based on findings from a series of investigations into the perceptual strategies used to identify word boundaries in speech, the proposed hierarchy consists of three levels or tiers. The top tier, comprising higher-level linguistic information arising from syntactic, semantic, and lexical cues, is sufficient for word identification when listening conditions are optimal. The middle tier, encompassing segmental cues arising from phonotactic constraints and acoustic-phonetic features, contributes when higher-level cues are insufficient. The lowest tier, composed of suprasegmental cues arising from prosodic variation, comes into play when both higher-level linguistic and mid-level segmental cues are ambiguous.

The lower-tier cues contributing to lexical segmentation are those marking syllable strength. In English, most word-initial syllables and single-syllable words are characterized as strong (Cutler and Carter, 1987). They contain full vowels (i.e., they are not reduced to schwa) and may also receive prosodic stress (i.e., they have relatively higher F0, greater intensity, and longer duration). By contrast, syllables occurring in other word positions are more likely to be weak. They contain reduced vowels and lack prosodic stress (Fear et al., 1995; Cutler and Carter, 1987). When higher-level cues are impoverished, listeners' error patterns suggest that they are using their statistical knowledge of these syllabic stress patterns to segment continuous speech into words (e.g., Cutler and Butterfield, 1992; Liss et al., 1998). Specifically, they tend to erroneously insert lexical boundaries immediately before strong syllables, and delete

lexical boundaries before weak syllables. This has been termed the Metrical Segmentation Strategy (MSS) hypothesis (Cutler and Norris, 1988).

*Individual variability in cue use*

One implication of this hierarchical model is that, in any given condition, listeners must be adept at shifting their perceptual strategies to take advantage of the cues that remain relatively intact. The flexibility with which listeners can achieve this varies widely, even among individuals with normal hearing. Studies of individual variability in tasks such as the identification of phonemes (Hazan and Rosen, 1991), the discrimination of nonsense syllables (Surprenant and Watson, 2001), the perception of lexical tones (Chandrasekaran et al., 2010), and the comprehension of dysarthric speech (Choe et al., 2012), have shown large inter-subject differences in sensitivity to speech cues. This variability is not wholly explained by differences in auditory acuity. In their analysis of individual variation in auditory tasks involving both speech and non-speech stimuli, Surprenant and Watson (2001) observed only a weak correlation between spectro-temporal resolution and speech processing ability. Significant correlations between auditory and visual speech identification scores led them to argue for a cognitive-perceptual, rather than auditory peripheral, account of differences in speech processing performance.

Differences in cue use are magnified when the speech signal is reduced or degraded. In a study addressing variability in consonant identification, Hazan and Rosen (1991) observed large inter-individual performance differences when cues to place of

articulation for plosives were degraded via neutralization of either the burst frequency or the formant transition information. Further, they found that the two methods of cue reduction affected individual listeners differently; some showed similar effects from both manipulations, while others were more severely affected by one or the other. The authors noted that an understanding of individual differences in perceptual weighting of cues would be especially important for the effective treatment of clinical populations who have reduced access to a large array of speech cues (Hazan and Rosen, 1991). One group for whom this is a particular concern is cochlear implant users.

*Cochlear implants and speech cues*

Cochlear implants (CIs) afford most deafened individuals some degree of auditory speech recognition, but they do not restore normal communication. Both spectral and temporal detail are degraded for multiple reasons, including signal processing limitations, electrical current spread, and neural degeneration (see Wilson and Dorman, 2008, for a review). Reduced spectrotemporal resolution results in sparse cues to consonant place of articulation, vowel formant location, and variations in the F0 and intensity of the talker's voice. This leads to difficulty performing segmental tasks such as phoneme and word recognition (Gifford et al., 2008), and suprasegmental tasks such as talker identification based on differences in F0 (Fu et al., 2004), discrimination of prosodic contours (Meister et al., 2011), and use of syllabic stress cues to identify word boundaries (Spitzer et al., 2009). These difficulties are magnified in adverse conditions such as noise (Nelson et al., 2003), reverberation (Poissant et al., 2006; Helms Tillery et al., 2012), and poor telephone reception (Fu and Galvin, 2006).

Cue redundancy in the speech signal affords CI users some benefit. Evidence for this is found in the importance of both F0 and intensity variation to the processing of prosodic information. In a study of CI users' sensitivity to modulations in F0 and intensity in synthetically manipulated words, Rogers et al. (2006) found that combining the two cues led to a significant reduction (i.e., improvement) in listeners' thresholds for detecting the modulation, compared to thresholds for either cue alone. They also noted that sensitivity to combined F0/intensity variation was significantly correlated with word recognition. Brown and Bacon (2009) reported that CI users with residual acoustic hearing demonstrated improved sentence recognition when they combined the electric stimulation from their device with an acoustically-presented tone modulated to track both the amplitude envelope and F0 variation in target speech.

Further evidence for the usefulness of even degraded F0 cues is seen in the detrimental effects of reduced F0 variation. Spitzer et al. (2009) observed that implant users, whose acuity to F0 variation is relatively weak, still relied on F0 cues to perform lexical segmentation tasks. When the F0 contours in their stimuli were flattened, their CI participants demonstrated a decline in the use of syllable stress cues to segment speech. These findings suggest that, when segmental cues are impoverished, as in CI-processed speech, listeners do shift their cue weighting schemes toward a reliance on suprasegmental cues, even though these are also degraded. The degree to which individual CI users vary in their perceptual weighting of segmental and suprasegmental cues remains unclear, however.

Few reports in the CI literature address the question of differences in cue use across individual listeners. One group has conducted a series of studies examining variability in CI users' sensitivity to the acoustic features signaling prosodic contrasts (Peng et al., 2009, 2012). In naturally produced prosodic contours, F0 and intensity are co-modulated, such that as one rises or falls, the other does too. Both acoustic cues can thus provide prosodic information. To investigate CI users' sensitivity to these two acoustic markers, Peng and colleagues created words with prosodic contours marking either a statement or a yes-no question, then independently manipulated F0 and intensity variation to generate tokens in which the two cues conflicted. When they asked CI users to identify the words as questions or statements, they found that individual listeners were variably affected by the cue conflict. Some relied more heavily on F0 variation to make their determination (i.e., their judgments were less affected by the conflicting cues), while others tended to weight F0 and intensity more evenly (i.e., their judgments were more affected by the conflict). They noted that CI users showing decreased sensitivity to the intensity cue may not be making maximal use of all acoustic markers signaling prosodic contrasts in more natural conditions. A lack of sensitivity to suprasegmental cues would likely affect the flexibility with which CI users can shift their perceptual strategies to perform lexical segmentation tasks.

Studies of perceptual cue weighting in another population of listeners may shed additional light on this issue. Individuals with motor speech disorders such as dysarthria exhibit disrupted production of both the segmental and suprasegmental elements of speech. Listeners tasked with deciphering dysarthric speech thus must adapt their perceptual strategies to cope with cues that are degraded at both levels. In this case, the

stages of the perceptual hierarchy described by Mattys (2005) are not as clear-cut, and individual differences in cue weighting schemes become apparent.

In a study of individual variation in the perception of dysarthric speech, Choe et al. (2012) analyzed listener's transcriptions of phrases produced by speakers with dysarthria. The phrases, consisting of low semantic context and strong-weak or weak-strong syllable patterns, were designed to examine listeners' relative reliance on lower-tier phonemic and prosodic information to identify words. When phonemic cues are insufficient for word identification, listeners are predicted to resort to their knowledge of English stress patterns to locate lexical boundaries. In this case, a reliance on syllable stress is characterized by a greater proportion of (a) word boundary insertions before strong syllables (e.g., the target word *advance* was transcribed as "and then") and (b) word boundary deletions before weak syllables (e.g., the target words *frame her* were transcribed as "framer")[2]. On the other hand, when phonemic information is sufficient for word identification, error distributions are more even. The listener transcriptions analyzed by Choe et al. (2012) showed evidence for individual variability in perceptual weighting schemes. Some demonstrated higher phoneme accuracy and more even error distributions, while others showed lower phoneme accuracy and a higher proportion of word boundary errors, reflecting a reliance on syllable stress cues. As such, these findings are additional support for the notion that CI users, whose access to both sets of cues is degraded, may also demonstrate individual differences in cue weighting.

[2] Stressed syllables are indicated by underlining.

Peng and colleagues (2012) argued that variability in CI users' sensitivity to speech cues should be a consideration in the provision of perceptual training. However, the perceptual learning accomplished by adult CI users often is not guided by formal training. While several at-home computer-based auditory training programs are available to CI users, most do not allow for any comparative assessment of a listener's sensitivity to segmental and suprasegmental cues, and most do not provide focused practice with suprasegmental elements of speech (Appendix A).

Despite (or perhaps because of) a lack of clinic-based perceptual training for CI users, the mechanisms involved in the perceptual learning of CI-processed speech have received much attention from the research community. Many groups who study the perceptual learning of degraded speech employ vocoding as a way to simulate CI stimulation (e.g., Rosen et al., 1999; Loebach et al., 2008; Li et al., 2009; Krull et al., 2012). This allows examination of the effects of reduced spectral detail, and avoids potential confounds such as neural degeneration, frequency-place mismatch, and reduced dynamic range that often occur in actual CI users (Loebach et al., 2010). While acknowledged to have limitations, simulations are generally accepted as a valuable tool in the effort to gain insight into the perceptual learning of CI-processed speech.

*Perceptual learning of degraded speech*

Perceptual learning has been defined as changes in a perceptual system that occur in response to environmental influences and persist over time (Goldstone, 1998). Auditory perceptual learning is influenced by such factors as the duration of training

(e.g., Watson, 1980, 1991), training task (Davis et al., 2005), extent and type of feedback (e.g., Davis et al., 2005; Loebach et al., 2010, Borrie et al., 2012a, b), stimulus materials (e.g., Davis et al., 2005; Hervais-Adelman et al., 2008; Shafiro et al., 2012), and stimulus variability (e.g., Perrachione et al., 2011). Findings relevant to the perceptual learning of degraded speech are summarized briefly below.

*Duration of training.* Watson (1991) noted that the amount of practice needed by listeners to discriminate or identify unfamiliar complex waveforms can range from hours to months. He suggested that the learning of CI-processed speech may follow a similar time course, although cognitive–perceptual factors could play a mitigating role. Indeed, a study by Stacey and Summerfield (2008) showed that performance on phoneme, word, and sentence tasks improved significantly after an hour of training with vocoded stimuli. In an investigation of the perceptual learning of noise-vocoded sentences, Loebach et al. (2010) found that mean sentence recognition scores improved by about 20 percentage points over the course of 130 trials. Davis et al. (2005) observed gains of about 35 percentage points after 30 trials with high-context sentences. These results suggest that perceptual learning of both sublexical speech units and context-rich spectrally degraded sentences can occur over relatively short periods of time, at least with simulated cochlear implant processing.

*Training task.* In an investigation of the effects of passive listening, Davis et al. (2005) asked subjects to listen (without responding) to sentences presented in an iterative vocoded-clear-vocoded format, then tested their recognition of vocoded sentences. Mean post-training accuracy was approximately 30 percentage points higher than pre-training

performance. They suggested that mere exposure to vocoded speech may be sufficient for perceptual learning to occur.

*Feedback.* Several studies have shown that feedback provided during training significantly improves performance (Davis et al., 2005; Hervais-Adelman et al., 2008; Wayne and Johnsrude, 2012; Loebach et al., 2008, 2010). Davis and colleagues (2005) reported that, while the pre-test to post-test gains achieved with training are about the same with or without feedback, overall accuracy is significantly higher with feedback. They also observed that, as long as it allows listeners to compare the degraded stimulus with a clear representation of the target, feedback can take a variety of forms (Hervais-Adelman et al., 2008). Clear speech, text, and lipreading cues all provide similar amounts of benefit (Davis et al., 2005; Wayne and Johnsrude, 2012). In a study of the role of feedback type in learning, Loebach et al. (2010) observed that text feedback was most effective when it was presented simultaneously with an auditory repetition of the degraded stimulus. They suggested that the act of reading along with the auditory signal may facilitate the development of a cognitive-perceptual link between the degraded acoustic signal and a clear mental representation of the target.

*Training specificity.* Most studies of the perceptual learning of CI-processed speech have focused on the effects of phoneme-, word-, or sentence-based training. Findings from several of these studies indicate that learning tends to be specifically tied to the types of stimuli used. Materials that focus on acoustic-phonetic features lead to greater gains in sentence intelligibility (Loebach et al., 2008). The converse has also been observed: that training with sentences may lead to improved consonant identification (Fu

et al., 2005). However, training on segmental cues does not necessarily generalize to suprasegmental stimuli. Zhang et al. (2012) found that training CI users on phonetic discrimination did not transfer to talker identification or emotion recognition. Krull et al. (2012) have done one of the few studies of suprasegmental training. They found that talker identification training generalized to tasks involving speech recognition, although results were not consistent across different CI users. While these findings suggest that training on suprasegmental cues may facilitate learning of CI-processed speech, the role of individual variability, as well as the effects of training on the perceptual weighting of segmental and suprasegmental cues, remain unclear.

Studies of the perception of dysarthric speech may lend some insight into the latter question. Borrie et al. (2012a, b) reported that perceptual strategies used in the lexical segmentation of dysarthric speech differ, depending on the level of feedback and the type of training. They found that listeners whose training involved simultaneous presentation of auditory and written (text) representations of the stimuli placed more weight on suprasegmental cues when making lexical decisions, while listeners who received no written feedback (or feedback of any form) tended to weight segmental cues (Borrie et al., 2012a). In a follow-up study, they observed that training materials that emphasized syllabic stress resulted in perceptual weighting of suprasegmental cues, regardless of the degree of feedback (Borrie et al., 2012b). Together, these findings suggest that when both segmental and suprasegmental cues are degraded, listeners' perceptual strategies may be influenced both by the focus of training and the degree of feedback.

*Stimulus variability.* Training paradigms that expose listeners to multiple exemplars of the target are thought to promote generalization to novel stimuli (Samuel and Kraljic, 2009). However, stimulus variability can also slow the rate of learning and reduce overall accuracy (Clopper and Pisoni, 2004). The effects of stimulus variability on learning and generalization are qualified by individual differences in perceptual sensitivity. In a study of the effects of talker variability on lexical tone learning in normal-hearing listeners, Perrachione et al. (2011) found that individuals with poorer pre-training pitch contour discrimination learned tone contrasts more slowly, were less accurate, and demonstrated less generalization than individuals with higher pre-training pitch discrimination. When the target talkers were presented in blocks to reduce trial-by-trial variability, the lower-performing listeners showed improved learning and generalization, while the higher-performing listeners showed performance levels similar to those in the unblocked condition. The authors concluded that a training paradigm that provides high overall stimulus variability in a blocked design that limits trial-by-trial uncertainty may facilitate perceptual learning and generalization for listeners with a range of individual abilities.

Taken together, results from these studies indicate that perceptual learning of degraded speech can occur over relatively short periods of time. Segmental-level learning can generalize to sentence-level tasks, but does not carry over to suprasegmental tasks. Reports are conflicting as to whether suprasegmental training generalizes to other speech tasks. In some contexts, passive exposure to speech stimuli may be sufficient to promote learning. Stimulus variability can facilitate learning and generalization when

similar stimuli are blocked together. Finally, higher overall accuracy is achieved when feedback includes an unambiguous representation of the target.

*Summary and present study*

In adverse conditions, when higher level lexical and segmental information is impoverished, listeners can resort to relatively robust suprasegmental information to parse speech. Cochlear implant users, however, face degradation at both the segmental *and* suprasegmental levels. For these listeners, the ability to use both types of cues, despite their degradation, is critical for speech perception. While segmental training has been shown to improve speech intelligibility, little is known about the potential benefits of suprasegmental training. In addition, the effects of individual variability in cue weighting on perceptual learning are unclear.

The present study used simulated cochlear implant stimulation to examine the effects of training specificity and individual differences in cue weighting on the perceptual learning of degraded speech. The study design involved a series of alternating testing and training procedures that took place over two sessions. During Session One, participants took a series of baseline tests, completed a sentence transcription task designed to increase their familiarity with the degraded signal, and then took a series of follow-up tests. During Session Two, participants received targeted training with either segmental or suprasegmental cues while receiving passive exposure to the other cue type, then took a final series of tests. Both the Familiarization and Targeted Training tasks were designed to maximize opportunities for perceptual learning by leveraging

previously reported benefits of high-context materials, passive exposure, training specificity, stimulus blocking, and feedback.

Analyses included within-subject and between-group comparisons to address the following questions:

1) Can targeted training increase acuity to degraded segmental or suprasegmental cues?
2) Does increased acuity to segmental or suprasegmental cues generalize to improvements on other speech tasks?
3) Does targeted training lead to the increased use of these cues to parse degraded speech?
4) Is passive exposure sufficient to increase acuity to, and use of, these cues, or is active attention necessary?
5) How do individual differences in perceptual cue weighting affect training outcomes?

Findings were examined with an eye toward profiling differences in perceptual learning and exploring how these profiles might inform clinical decisions about targets for aural rehabilitation.

## Method

### *Participants*

Eighty-six native speakers of North American English (17 males) from Arizona State University (ASU) and the surrounding area participated. Six did not complete the experiment due to unforeseen scheduling problems. Participants were between the ages of 18 and 36 (mean = 22.9, SD = 3.95), had self-reported normal or corrected-to-normal vision, and underwent audiological screening to verify pure-tone thresholds of 20 dB HL or better at octave and half-octave frequencies from 125 to 8000 Hz (ANSI, 2004). All reported having no prior experience with CI simulations. The study was approved by the Institutional Review Board for Human Subjects Research at ASU (Appendix B). All subjects provided their written informed consent to participate, and received hourly compensation for their time.

### *Speech materials*

Six stimulus sets were used during testing; two of them were also used during Familiarization and Targeted Training. Each targeted a specific speech skill, including prosody and phoneme discrimination, vowel and consonant recognition, and phrase and sentence transcription. The stimuli are described briefly below.

*Word triplets.* A corpus of 260 rhyming word triplets was developed to target discrimination of either segmental or suprasegmental differences across words. Each triplet contained one word with a contrasting prosodic contour, and one with a contrasting initial phoneme (e.g., “Bill. Bill? Fill.”). Two hundred triplets were used during training, and 20 were used during each test. The stimuli were designed to afford targeted

practice with one of the contrasts, while providing passive exposure to the other contrast. This allowed examination of whether active attention is necessary to increase sensitivity to segmental or suprasegmental cues. Descriptions of the stimulus design parameters and procedures for validating the prosodic contours are in Appendix C, along with a list of the words used to form the triplets.

*Vowels and consonants.* Ten /h-vowel-d/ words (heed, hid, hayed, head, had, hawed, HUD, hoed, hood, who'd) and twenty /a-consonant-a/ disyllables containing the phonemes /p, b, t, d, k, g, f, v, s, z, ch, j, sh, th, l, r, m, n, w, h/ were used to test vowel and consonant recognition. These stimulus sets also allowed examination of whether increased sensitivity to segmental cues would generalize to improvement on other segmental tasks.

*Question-statement pairs.* Sixty short sentences, produced either as a question or as a statement, were generated to test sentence-level prosody discrimination (e.g., "She baked a small cake." vs. "She baked a small cake?"). Twenty sentences were used in each test. These stimuli were used to assess whether increased sensitivity to suprasegmental cues would lead to improvement on another suprasegmental task. A list of the sentences and a description of the procedures for validating the prosodic contours are in Appendix D.

*High context sentences.* One hundred forty sentences from the CUNY corpus (Boothroyd et al, 1988) were used to target sentence transcription. Sentences vary in length and contain between two and nine key words. Characterized as having high semantic and syntactic context and produced with marked variations in prosodic contour, the CUNY sentences provided exposure to degraded segmental and suprasegmental

speech cues during the Familiarization task. Eighty sentences were used during Familiarization, and 20 were used for each test.

*Low-context phrases.* One hundred twenty phrases with low inter-word predictability (Liss et al., 1998) were used to assess 1) the ability to identify words in low-context phrases and, 2) the perceptual strategies used when decoding these stimuli. Each phrase contained three to five words, and consisted of six syllables with alternating stress patterns. In half the phrases, syllabic stress followed a trochaic (i.e., strong-weak) pattern (e.g., *soon the men were asking*), and in the other half it followed an iambic (i.e., weak-strong) pattern (e.g., *amend the slower page*). The phrases are characterized as syntactically plausible but semantically anomalous. When they are degraded (e.g., by vocoding), listeners must parse them using their 1) higher-level knowledge of the lexicon, and 2) ability to extract ambiguous lower-level phonemic and prosodic cues (Liss et al., 1998; Spitzer et al., 2009). Analysis of the types and locations of lexical boundary errors listeners make provides information about the degree to which they use syllabic strength to identify word boundaries in the phrases. For example, a higher proportion of lexical boundary insertions before strong syllables would indicate greater reliance on syllable strength cues. On the other hand, a more even error distribution would suggest greater use of segmental cues for word identification.

*Audio recordings*

Each stimulus set was recorded digitally (sample rate: 44.1 kHz, amplitude resolution: 32 bit) in a double-walled sound-attenuated booth using an AKG C2000B microphone and Audacity sound editing software. Half of each set was produced by a

male talker (mean F0 = 131 Hz) and half by a female (mean F0 = 212 Hz), both native speakers of North American English. Recordings in each stimulus set were equated with respect to RMS, and peak values were normalized.

*Signal processing*

Equated and normalized stimuli were processed through a six-channel noise-excited vocoder whose carrier bands were shifted up in frequency relative to the analysis bands. Vocoding was accomplished via the following steps: analysis filtering of the speech signal into six contiguous frequency bands; half-wave rectification and low-pass filtering ($6^{th}$ order Butterworth) to extract the amplitude envelope of each band; modulation of a noise carrier by each extracted envelope; output filtering of the noise carriers; and summation of the modulated output bands to simulate CI-processed speech. The low-pass cut-off frequency of the envelope filter was 400 Hz or half the bandwidth of the input band, whichever was lower. The logarithmically spaced cut-off frequencies of the analysis bands ranged from 100 Hz to 6000 Hz, while those of the carrier bands ranged from 226 Hz to 9156 Hz. This simulated approximately a 3mm basalward frequency shift in the cochlea (Greenwood, 1990), which has been shown to degrade speech cues while avoiding floor effects (Dorman et al., 1997a; Fu and Shannon, 1999a).

Digital processing of stimuli was accomplished via custom Python scripts. Stimuli were output through an Echo Gina3G soundcard, and adjusted to an overall level of 70 dB SPL with a Tucker-Davis PA5 programmable attenuator before they were presented monaurally to listeners.

*Procedure*

The sessions, both lasting about two hours, were completed on consecutive days. Figure 1 provides a schematic representation of the conditions in each session. For the first session, the testing and familiarization procedures were identical for all participants. For the second session, participants were assigned in a blocked manner to receive targeted training with one of two Training Strategies (phonemic or prosodic) and one of two Signal Types (degraded or clear). Training with the clear stimuli was intended as a control for the training with the degraded stimuli. Forty participants received training with phonemic cues, and 40 received training with prosodic cues. Within each Training Strategy, half of the participants were trained with the degraded signal, and half with the clear signal.

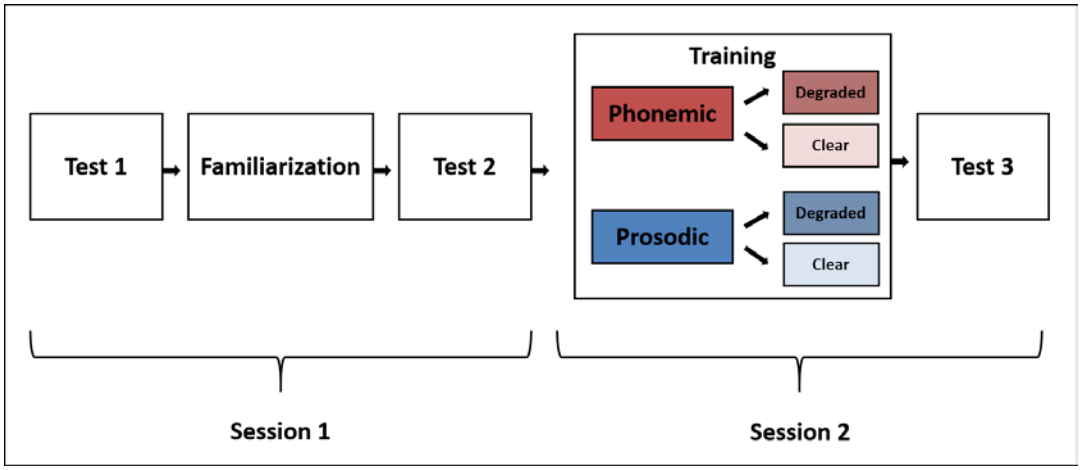



*Figure 1*. Schematic representation of experimental procedures and conditions. Phonemic training is indicated in red, prosodic in blue. Signal type is shown by shading: darker shades represent training with the degraded signal; lighter shades represent the clear signal.

Sessions were conducted in a double-walled sound-attenuated booth, in which participants wore AKG K271 MKII headphones and sat facing a computer monitor. Stimulus presentation and data collection were controlled via Presentation, a commercial software package for behavioral experiments (version 17.1, www.neurobs.com). Each experimental task was preceded by written instructions (Appendix E), and each auditory stimulus was preceded by a visual prompt to listen. Instructions and prompts were presented as text on the monitor. Participants typed their responses on a keyboard, and were encouraged to guess when unsure of the answer. Experimental tasks are described below.

*Tests:* Participants were given a battery of seven speech tests at three time points: Baseline (Test 1), after Familiarization (Test 2), and after Training (Test 3). All test stimuli were degraded. Prior to each of the Baseline tests, listeners completed 10 practice items presented with clear (non-degraded) speech. None of the practice items were used during testing. Test order and individual test stimuli were randomized across listeners. Listeners did not receive feedback during practice or testing. Except for the low-context phrase transcription test, stimulus tokens were presented only once. Because the phrases were particularly difficult to transcribe, forty of the listeners were allowed to hear each low-context phrase up to 3 times. This increased their attempts at word identification and allowed for a more robust analysis of their lexical boundary errors.[3]

[3] To determine whether the opportunity to repeat the phrase stimuli affected performance on other measures, independent samples t-tests were conducted on data from the "Repeats" group and the "No Repeats" group for each of the seven test measures. On the phrase transcription task itself, accuracy was significantly higher for the "Repeats" group. Performance on all the other test measures was not significantly different across the two groups ($p > .05$). In addition, results from $\chi^2$ goodness of fit analyses of the lexical boundary error distributions from the two groups indicated that they were not significantly different. Thus the data from the two groups were pooled for all analyses.

*Familiarization.* Eighty high-context CUNY sentences were used for Familiarization, which occurred during Session One. Listeners transcribed each sentence after hearing a single degraded auditory presentation. They then received the answer, in the form of text accompanied by an auditory repetition of the degraded sentence. Sentences were presented in sets of 10, and blocked by talker (40 male; 40 female). Sentence set and talker order were randomized across listener. This task provided an opportunity for non-guided incidental learning of the degraded signal (i.e., it did not specifically target attention to either phonemic or prosodic cues). It was expected to result in overall increases in segmental acuity and corresponding decreases in reliance on suprasegmental cues for lexical segmentation, in accordance with the MSS hypothesis, which would allow room for the potential effects in Session Two of suprasegmental training. It was also intended to allow listeners to stabilize their own blend of perceptual strategies for decoding degraded speech. Familiarization lasted approximately 30-45 minutes.

*Targeted Training:* Two hundred rhyming triplets were used for Targeted Training, which occurred during Session Two. For each triplet, listeners selected the word containing either the different phoneme, or the different prosodic contour, depending on their Training Strategy assignment. After responding, listeners heard the triplet again while the words appeared on the monitor with the answer highlighted. The task was intended to expose listeners to both phonemic (i.e., segmental) and prosodic (i.e., suprasegmental) cues, while providing targeted practice with just one of the cue types. The training stimuli were degraded for half of the listeners in each Training Strategy group, and clear for the other half. Stimuli were blocked by talker (100 male,

100 female), but randomized within each talker set. Talker order was randomized across listener. Targeted Training lasted approximately 30-45 minutes.

*Scoring and Reliability*

Responses to the prosody, phoneme, vowel, consonant, and question/statement tasks were scored for percent correct. Transcripts of the CUNY sentences were scored by a trained rater for percent keywords correct.

Transcripts of the low-context phrases were coded by a trained rater for (a) the number of words correctly transcribed, and (b) the number, type, and location of lexical boundary errors (LBEs). Error type was coded as insertion (I) or deletion (D) of a syllable. Error location was coded as before a strong syllable (S) or before a weak syllable (W). Transcripts were thus scored for four possible error combinations: insertion of a word boundary before a strong syllable (IS) or before a weak syllable (IW); deletion of a word boundary before a strong syllable (DS) or before a weak syllable (DW). Errors were tallied for each listener, and were pooled in each training group.

Given that reliance on syllable stress for lexical segmentation is predicted to result in more word boundary insertions before strong syllables than weak (and more deletions before weak syllables than strong), the ratio of predicted to non-predicted errors for each error type indicates strength of adherence to the Metrical Segmentation Strategy (Spitzer et al., 2007). Calculations of the IS/IW and DW/DS ratios were thus made for each listener and each training group. Finally, the ratio of IS and DW errors to the total number of errors was calculated for each listener and group, to obtain an overall

measure of metrical segmentation. This last calculation has been termed the MSS ratio (Spitzer et al., 2007).

To obtain reliability estimates for the LBE coding, a second trained rater independently scored one third of the transcripts. There was a high degree of inter-rater consistency, as indicated by a Cronbach's alpha of .948. The raters' LBE distributions were also subjected to a $\chi^2$ goodness of fit test to determine whether they were drawn from the same sample. Results were not significant [$\chi^2$ (3) = .367, p = .947], indicating that the raters' LBE distributions were not different. Inter-rater scoring discrepancies were resolved or tokens were discarded. Less than one percent of the responses were discarded due to unresolvable discrepancies.

*Analyses*

The effects of Familiarization and Training were examined via group analyses and individual comparisons. Group Familiarization effects were assessed using repeated measures analyses of variance (ANOVAs) to compare Test 2 measures to their counterparts from Test 1. Group effects for Targeted Training were investigated by subjecting the data from Tests 2 and 3 to three-way mixed ANOVAs with Training Strategy and Signal Type as between-groups factors.

Individual differences were investigated by comparing the performance of selected exemplar listeners on a range of tasks. Given previous work showing individual variation in the contributions of both segmental and suprasegmental skills to the intelligibility of the low-context phrases (Choe et al., 2012), listeners' improvement on the phrase transcription task was treated as a reference point against which to

examine their changes in consonant and vowel recognition, and use of the Metrical Segmentation Strategy. Sentence transcription scores were also included as an indicator of their use of higher-level lexical cues. Listener profiles showing relative performance on these measures were used to gain insight into how initial abilities, along with the learning of these cues, may contribute to the ability to decode degraded speech.

## Results

### *Familiarization effects*

To confirm that the Familiarization task[4] facilitated incidental learning of degraded speech cues, performance on Test 2 was compared to that on Test 1. Group effects were examined for each speech measure, including lexical boundary error distributions. Individual analyses were aimed at examining clusters of speech skills for listeners showing different degrees of gain on the phrase transcription task.

#### *Group analyses*

*Speech measures.* Figure 2 shows mean percent correct on the speech measures from Tests 1 and 2. Repeated-measures ANOVAs indicated that Test 2 scores were significantly higher than those from Test 1 for all seven measures (Table 1).

*Lexical boundary errors*. The pooled LBEs from Test 1 were compared to those from Test 2. Results from a $\chi^2$ goodness of fit analysis, conducted to examine the null hypothesis that the two LBE distributions were drawn from the same sample, were significant [$\chi^2$ (3) = .116.63, p < .001]. This indicated that the ratios of predicted to non-predicted errors on Test 1 were different from those on Test 2 (Table 2). Familiarization resulted in an overall reduction in reliance on the Metrical Segmentation Strategy to

[4] Sentences in the Familiarization task were presented in blocks of ten. Mean percent keywords correct over the entire task was 72%. Repeated measures ANOVAs performed on the sentence blocks revealed a significant main effect of block [$F(7, 79) = 5.427$, $p < .001$, $\eta_p^2 = .064$], indicating that accuracy improved over time. *Post hoc* Bonferroni tests revealed significantly higher scores on the last block (76%) than on the first (69%) [$p = .024$]. In addition, scores on the last three blocks were not significantly different from each other ($p = 1.0$), indicating that an asymptote in performance was reached.

identify words in the low-context phrases, supported by a reduction in the ratio of predicted to non-predicted LBEs with an overall increase in intelligibility (figure 3).

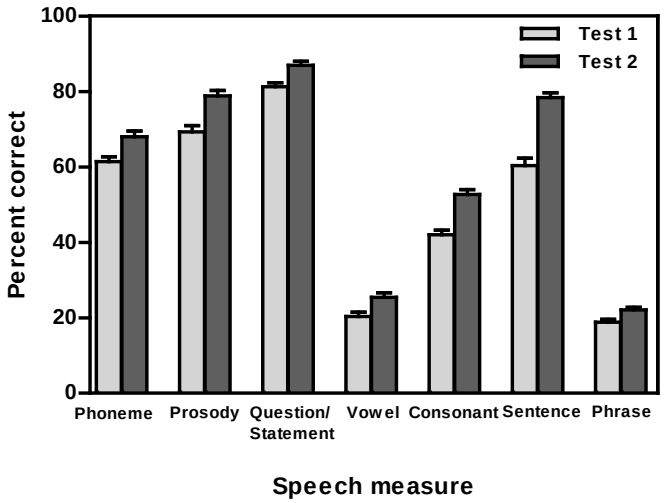



*Figure 2.* Mean percent correct on Test 1 and 2 speech measures. Error bars represent standard error.

Table 1.
*Repeated-measures analyses of variance for Tests 1 and 2*

| Test | $F(1, 79)$ | $p$ | $\eta_p^2$ |
|---|---|---|---|
| Phoneme discrimination | 19.82 | < .001* | .201 |
| Prosody discrimination | 39.16 | < .001* | .331 |
| Question/statement discrimination | 34.09 | < .001* | .301 |
| Vowel recognition | 14.90 | < .001* | .159 |
| Consonant recognition | 90.65 | < .001* | .534 |
| Sentence transcription | 174.60 | < .001* | .688 |
| Phrase transcription | 21.25 | < .001* | .212 |

*Note.* A Bonferroni correction for seven comparisons was applied to each test.

Table 2.
*Distributions of pooled LBEs (and percent of total occurrence) in the phrase transcription task in Tests 1 and 2.*

| | IS | IW | DS | DW |
|---|---|---|---|---|
| Test 1 | 1477 (45%) | 816 (25%) | 404 (12%) | 551 (17%) |
| Test 2 | 1484 (44%) | 1000 (30%) | 494 (15%) | 390 (12%) |

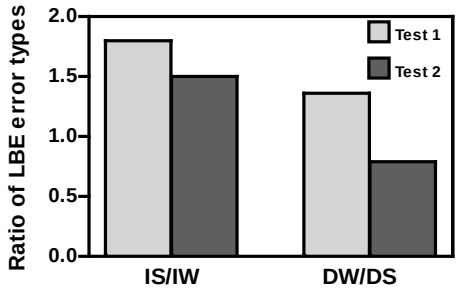



*Figure 3.* Ratio of predicted to non-predicted LBEs for insertion and deletion errors on Tests 1 and 2.

*Individual analyses*

While average performance improved after Familiarization, there was considerable inter-individual variability in the amount of gain, and some listeners even showed declines. This was particularly evident in the phrase transcription task. Figure 4, panel A, shows individual Test 2 phrase scores plotted against those from Test 1. Points above the diagonal line represent an increase after Familiarization; those below a decrease. Distance from the line indicates degree of change. Listeners with the highest scores on Test 1 tended to show a drop on Test 2, while the converse occurred for those with the lowest Test 1 scores. The largest increases in accuracy tended to occur for listeners in the mid-range. Similar trends were apparent in the ratio of predicted to non-predicted lexical boundary insertion errors (IS/IW ratio, panel B). Most listeners showed improvements on consonant recognition (panel C) and sentence transcription accuracy (panel E). Overall performance on the vowel recognition task (panel D) was low, and several listeners showed little improvement after Familiarization.

Pearson correlation analyses, conducted to determine whether improvement on phrase accuracy was associated with improvement on other tasks, revealed a significant correlation between phrases and sentences ($r$ (80) = .303, $p \leq .006$). All other associations were not significant ($p \geq .069$).

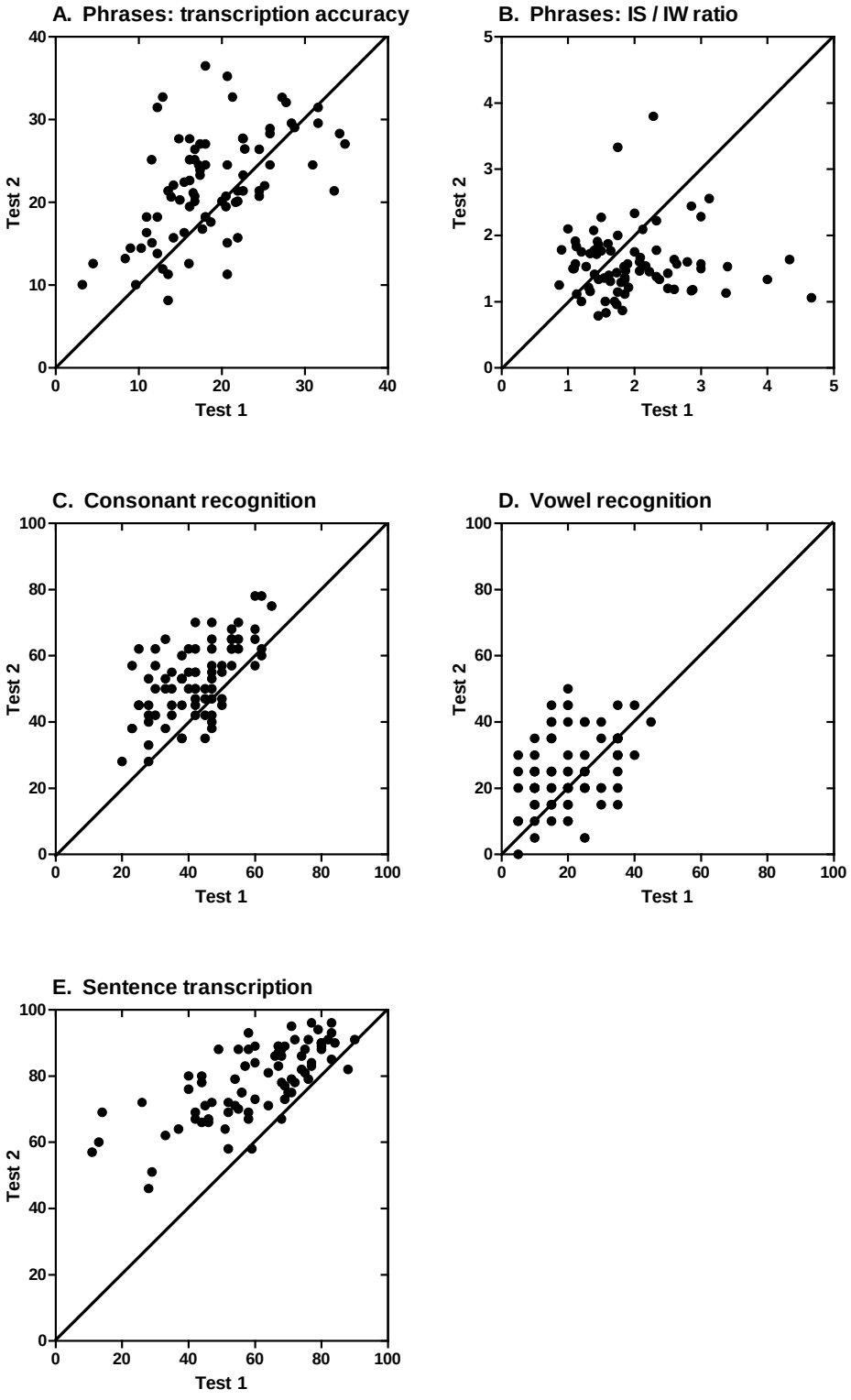



*Figure 4.* Scatterplots of Test 2 scores vs. those on Test 1. Note: not all plots are on the same scale.

To further examine individual differences, the test scores of three pairs of exemplar listeners were selected for qualitative comparison (figure 5). One pair represented lower phrase accuracy on Test 1 (participants 29 and 82), one represented mid-range performance (participants 25 and 69), and one represented higher Test 1 phrase accuracy (participants 45 and 94). In each pair, one listener showed improved phrase accuracy on Test 2, and the other did not. The pairing options for the high-performing listeners were limited (participant 45, the closest match to 94, showed only small gains in phrase accuracy after Familiarization).

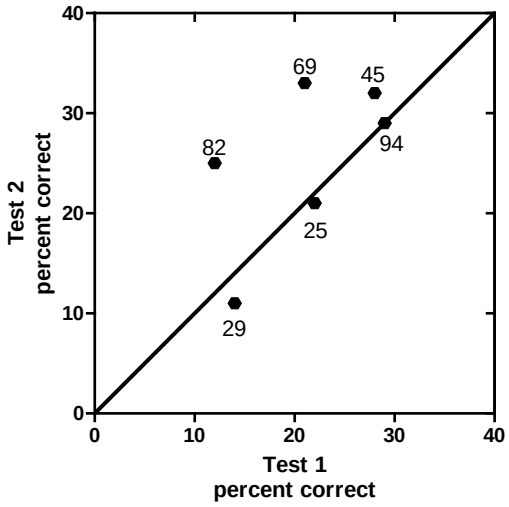



*Figure 5.* Test 2 phrase transcription scores for 3 pairs of listeners who showed lower (29, 82), mid-range (25, 69) and higher (45, 94) performance on Test 1.

Figure 6 illustrates the different clusters of skills shown by the six listeners. Panel A show baseline performance on the measures of interest. Black bars represents IS/IW ratio (left Y-axis). Consonant, vowel and sentence scores are represented by medium gray, light gray, and white bars, respectively (right Y-axis).

The singularity of each listener's baseline profile is clear (panel A). The two exemplars with low baseline phrase accuracy (participants 29 and 82) were similar in their low consonant, vowel, and sentence scores, but differed in IS/IW ratio. The listeners with mid-range phrase accuracy (participants 25 and 69) were similar only in their very low scores on vowel recognition. The listeners with higher baseline phrase accuracy (participants 94 and 45) differed across all measures. Across all the exemplars, one measure showing a fairly consistent trend was performance on sentences, which tended to reflect that on phrases (e.g., the pair with "Low" phrase accuracy showed lower performance on sentences, while the pair with "High" phrase accuracy showed higher performance on sentences).

In panel B of figure 6, the stacked purple bars represent the degree and direction of change on each measure after Familiarization. The listeners whose phrase transcription accuracy *did not* improve (participants 29, 25, and 94) showed declines or only small gains in most measures. Those whose phrase accuracy *did* improve (participants 82, 69, and 45) showed relatively large gains in at least one skill area: Listener 82 showed a large increase in IS/IW ratio, while listener 69 increased in vowel accuracy. Both showed large increases in sentence accuracy. Listener 45 showed a large increase in consonant accuracy, and was already near ceiling on sentence transcription.

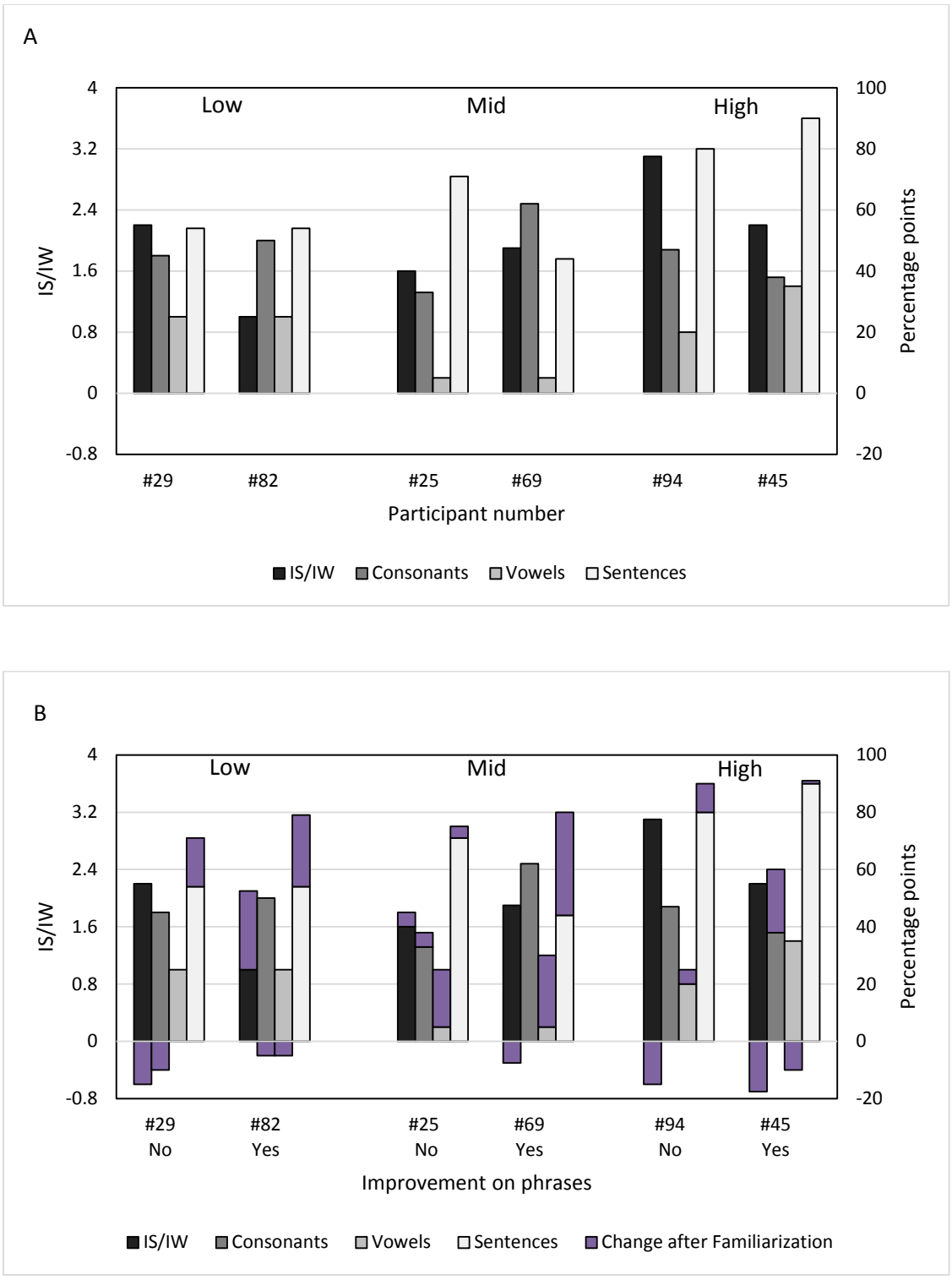



*Figure 6.* Speech skill profiles of three pairs of exemplar listeners who showed low-, mid- and high Test 1 phrase accuracy (panel A), and either improved or did not improve after Familiarization (panel B).

*Summary and discussion of Familiarization effects.* On average, performance on all seven speech tests improved after Familiarization, and overall use of the Metrical Segmentation Strategy decreased. There was considerable variability in baseline performance and degree of improvement on all measures, including phrase transcription accuracy. Those who were higher performers on phrase transcription before Familiarization showed less improvement, and in some cases a decline. The skills demonstrated by the higher-performing exemplars suggests that these listeners may already have been efficient users of multiple cues, limiting any further benefit of non-guided Familiarization. Those who were lower performers on phrase transcription before Familiarization generally showed only small gains. Examination of the scores of the low-performing exemplars indicated that they also had relatively low segmental and lexical scores, and demonstrated inconsistent improvement in these areas after Familiarization. For these listeners, the non-specific nature of the Familiarization task may have been insufficient for the learning of these cues. On the other hand, many of the mid-range performers showed large gains, indicating that practice with high-context sentences led to an improved ability to use both lexical and sublexical cues for the transcription of low-context phrases. Finally, the variability in skill level within each pair of exemplars suggests that similar levels of performance may be reached via different combinations of skills.

## *Targeted Training effects*

To examine the effects of specific training target on the perceptual learning of degraded speech, performance on Test 3 was compared to that on Test 2. [5] Scores on the training task itself were also reported. Group effects were assessed for each speech measure, including lexical boundary errors. Individual analyses were aimed at comparing the gains made by listeners receiving prosodic and phonemic training, given different pre-training levels of performance.

### *Group analyses*

*Performance on the training task.* Accuracy on the training task varied by training group (Table 3). A two-way ANOVA revealed significant main effects of Training Strategy [$F(1, 76) = 8.56$, $p = .005$, $\eta_p^2 = .101$] and Signal Type [$F(1, 76) = 24.88$, $p < .001$, $\eta_p^2 = .247$]. The interaction was not significant [$F(1, 76) = 1.05$, $p > .05$)]. *Post hoc* Bonferroni tests showed that mean performance on the Phonemic training task (90%) was significantly higher than the Prosodic training task (83%) [$F(1, 78) = 6.55$, $p = .012$, $\eta_p^2 = .077$]. Further, responses to the clear stimuli (93%) were significantly more accurate than those to the degraded stimuli (81%) [$F(1, 78) = 22.66$, $p < .001$, $\eta_p^2 = .225$]. [6]

[5] To confirm that there were no differences across training group before the start of Targeted Training, the data from each Test 2 speech measure were subjected to two-way ANOVAs with two levels of Training Strategy (phonemic, prosodic) and two levels of Signal Type (clear, degraded). All four groups commenced training with equivalent levels of performance on all measures [all $F(1,76) \leq 1.95$, $p \geq .166$].

[6] There were violations of the assumptions of normality and homogeneity of variance in the Training data set. Analyses were re-run with rationalized arcsine transformed data (Studebaker, 1985). Results were

Table 3.
*Mean percent correct training scores (and standard deviations) for each group.*

| | Signal Type | |
|---|---|---|
| Training Strategy | Degraded | Clear |
| Prosodic | 79 (15.28) | 88 (12.00) |
| Phonemic | 83 (6.67) | 97 (2.28) |

*Speech measures.* Table 4 provides a breakdown of scores by group for each speech measure in Tests 2 and 3. Scores from each of the measures were subjected to three-way mixed ANOVAs with two levels of Test Session (Test 2, Test 3), two levels of Training Strategy (phonemic, prosodic), and two levels of Signal Type (degraded, clear). *Post hoc* t-tests with Bonferroni corrections were used to follow up main effects and interactions. Results of the group analyses are broken into 1) training-specific effects, 2) generalization to other tasks, and 3) lexical boundary errors.

similar: there were main effects of Signal Type [$F(1, 76) = 37.82$, $p < .001$, $\eta_p^2 = .332$] and Training Strategy [$F(1, 76) = 8.57$, $p = .005$, $\eta_p^2 = .101$].

Table 4.
*Mean percent correct scores (and standard deviations) on Test 2 and 3 measures for each training group.*

| | Prosodic Training | | | | Phonemic Training | | | |
|---|---|---|---|---|---|---|---|---|
| | Degraded | | Clear | | Degraded | | Clear | |
| Measure | Test 2 | Test 3 | Test 2 | Test 3 | Test 2 | Test 3 | Test 2 | Test 3 |
| Phoneme | 68 (13.13) | 68 (17.20) | 68 (14.35) | 75 (6.97) | 71 (15.15) | 77 (6.97) | 66 (12.76) | 70 (13.13) |
| Prosody | 80 (16.26) | 81 (17.67) | 77 (13.42) | 84 (10.89) | 80 (10.70) | 78 (11.18) | 79 (12.15) | 73 (12.93) |
| Q/S | 87 (9.69) | 90 (7.95) | 87 (10.30) | 88 (8.01) | 89 (9.85) | 88 (7.77) | 85 (6.96) | 88 (7.78) |
| Vowel | 26 (11.54) | 31 (8.72) | 25 (10.38) | 28 (10.68) | 26 (8.87) | 30 (14.48) | 25 (13.32) | 29 (11.48) |
| Consonant | 52 (10.25) | 57 (9.55) | 54 (9.20) | 54 (10.4) | 50 (11.68) | 55 (10.51) | 55 (13.79) | 55 (12.70) |
| Sentence | 79 (11.13) | 85 (10.17) | 78 (13.12) | 83 (13.40) | 79 (11.22) | 87 (9.64) | 78 (9.98) | 84 (8.71) |
| Phrase | 23 (7.51) | 30 (11.15) | 21 (6.22) | 27 (8.38) | 23 (6.19) | 30 (8.68) | 21 (5.84) | 25 (7.74) |

Note: Q/S = Question/statement

*Training-specific effects.* Improvement on the phoneme and prosody measures was specific to the type of training received.

Overall phoneme discrimination accuracy was significantly higher after training than before [$F(1, 76) = 6.199$, $p = .015$, $\eta_p^2 = .075$]. The main effects of Training Strategy and Signal Type were not significant [all $F(1, 76) \leq 0.25$, $p \geq .619$], but there was an interaction between the two [$F(1, 76) = 5.419$, $p = .023$, $\eta_p^2 = .067$]. For listeners hearing the degraded signal, those who received phonemic training scored significantly higher on the phoneme test than those who received prosodic training [$F(1, 76) = 3.99$, $p = .049$, $\eta_p^2 = .050$]. Performance by listeners hearing the clear signal was not different across Training Strategy [$F(1, 76) = 1.67$, $p = .200$].

This interaction was observed in the pooled data from the two test sessions, and as such is difficult to interpret. To further explore possible group differences, a two-way ANOVA was conducted with only the post-training (Test 3) data. Results were similar to those from the three-way ANOVA. The main effects of Training Strategy and Signal Type were not significant [all $F(1, 76) \leq .56$, $p \geq .454$], but there was a significant interaction [$F(1, 76) = 7.96$, $p = .006$, $\eta_p^2 = .095$]. For the listeners hearing the degraded signal, those receiving phonemic training scored higher on the phoneme test than those receiving prosodic training [$F(1, 76) = 6.383$, $p = .014$, $\eta_p^2 = .077$]. Performance by listeners hearing the clear signal was not significantly different across Training Strategy [$F(1, 76) = 2.14$, $p = .148$] (figure 7).

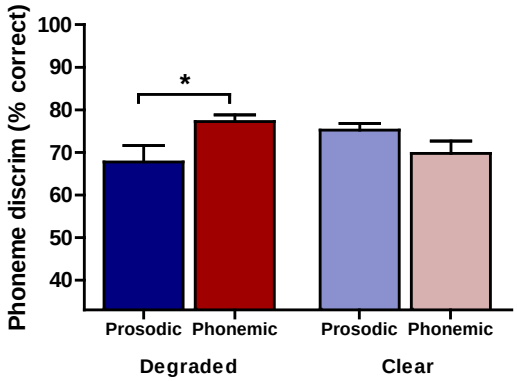



*Figure 7.* Phoneme discrimination scores on Test 3 as a function of training group assignment. Prosodic training is shown in blue; phonemic in red. The degraded signal is represented by darker shades, the clear signal by lighter shades. The Y-axis starts at chance (33%). Error bars represent standard error.

Prosody discrimination also showed specific training effects (figure 8). The main effects of Test Session, Training Strategy, and Signal Type were not significant [all $F(1, 76) \leq 1.58$, $p \geq .212$], but there was a significant interaction between Test Session and Training Strategy [$F(1, 76) = 6.622$, $p = .012$, $\eta_p^2 = .080$]. Listeners receiving prosodic training, regardless of signal type, scored significantly higher on the Test 3 prosody measure than those receiving phonemic training [$F(1, 76) = 5.62$, $p = .020$, $\eta_p^2 = .069$].

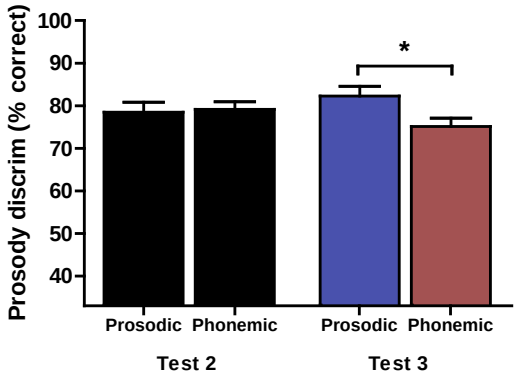



*Figure 8.* Prosody discrimination scores on Test 2 and Test 3 as a function of Training Strategy. Scores are collapsed across Signal Type (i.e., degraded / clear). The Y-axis starts at chance. Error bars represent standard error.

*Generalization to other tasks.* Scores on the other five speech tests increased after training. For all but the consonant measure, there were no significant group differences. Results are summarized for each test.

Performance on the question/statement discrimination task was near ceiling before training (87%), and improvements were small (2 percentage points). There were no significant effects of Training Strategy or Signal Type [all $F(1, 76) \leq 2.69$, $p \geq .105$], and no interactions were significant [all $F(1, 76) \leq 0.09$, $p \geq .116$].

Vowel recognition was significantly more accurate after training [$F(1,76) = 6.882$, $p = .011$, $\eta_p^2 = .083$]. On average, scores on Test 3 were five percentage points higher than those on Test 2. No other main effects were significant [all $F \geq 0.61$, $p \geq .436$], and no interactions were significant [all $F(1,76) \geq 0.90$, $p \geq .436$].

Consonant recognition gains were small but significant [$F(1,76) = 6.706$, $p = .012$, $\eta_p^2 = .081$]. On average, scores on Test 3 were 2.5 percentage points higher than those on Test 2. No other main effects were significant [all $F \geq 0.37$, $p \geq .547$]. There was a significant interaction between Test Session and Signal Type [$F(1,76) = 4.879$, $p = .03$, $\eta_p^2 = .06$]. Participants receiving training with the degraded signal made greater gains in consonant recognition than those receiving training with the clear signal [$F(1,76) = 11.51$, $p = .001$, $\eta_p^2 = .132$].

Sentence transcription was significantly more accurate after training [$F(1, 76) = 49.15$, $p < .001$, $\eta_p^2 = .393$]. On average, scores on Test 3 were six percentage points higher than those on Test 2. No other main effects were significant [all $F \geq 0.54$, $p \geq .463$], and no interactions were significant [all $F \geq 1.73$, $p \geq .192$].

Phrase transcription was also significantly more accurate after training [$F(1,76) = 78.03$, $p < .001$, $\eta_p^2 = .507$]. Mean performance on Test 3 was six percentage points higher than on Test 2. No other main effects were significant [all $F(1,76) \leq 2.66$, $p \geq .107$], and no interactions were significant [all $F \geq 0.90$, $p \geq .436$]. While not significant, the groups receiving training with the degraded signal did demonstrate greater gains than those receiving training with the clear signal.

*Lexical boundary errors*. Both of the groups receiving training with the degraded signal significantly increased their proportion of predicted to non-predicted lexical boundary insertion errors. This is evident in the results of within-group comparisons of the LBE distributions from Test 2 and Test 3. For each of the training groups, the null hypothesis that their pre- and post-training LBE distributions were drawn from the same sample was tested with a $\chi^2$ goodness of fit analysis. Results indicate that the groups who heard the degraded signal showed significant shifts in their error distributions after training, while the groups who heard the clear signal did not (Table 5). This is illustrated in figure 9, which shows greater shifts in the ratios of predicted to non-predicted insertion errors for the groups receiving training with the degraded signal (dark blue and red bars) compared to those receiving training with the clear signal (light blue and red bars). Similar trends occurred with deletion errors (not shown).

Table 5.
*Pooled LBE distributions (including percent of total occurrence of LBEs) and $\chi^2$ goodness of fit analyses (d.f.=3) for Tests 2 and 3 for each training group.*

| | | IS | | IW | | DS | | DW | | $\chi^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Test 2 | Test 3 | Test 2 | Test 3 | Test 2 | Test 3 | Test 2 | Test 3 | |
| Prosodic | Degraded | 347 (44%) | 250 (48%) | 252 (32%) | 131 (25%) | 111 (14%) | 72 (14%) | 77 (10%) | 68 (13%) | 23.77* |
| | Clear | 373 (44%) | 283 (44%) | 243 (28%) | 174 (27%) | 129 (15%) | 104 (16%) | 112 (13%) | 79 (12%) | 1.64 |
| Phonemic | Degraded | 363 (43%) | 282 (49%) | 250 (30%) | 150 (26%) | 128 (15%) | 85 (15%) | 106 (12%) | 59 (10%) | 14.71* |
| | Clear | 401 (46%) | 281 (47%) | 255 (29%) | 159 (26%) | 126 (14%) | 94 (16%) | 95 (11%) | 69 (11%) | 3.71 |

*$p < .01$.

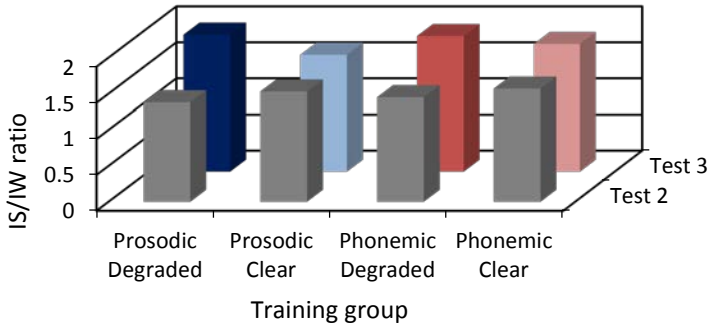



*Figure 9.* Ratio of predicted to non-predicted lexical boundary insertion errors made by each training group in Tests 2 and 3. Test 2 ratios are represented by the front row of gray bars; Test 3 ratios are shown by the back row of color-coded bars. Blue bars correspond to prosodic training; red to phonemic. Darker shades represent training with the degraded signal; lighter shades represent training with the clear signal.

*Individual analyses*

Examination of individual variability was limited to the groups receiving training with the degraded signal, given that significant group effects for two critical measures, consonant recognition and IS/IW ratio, were observed only for these listeners. Figure 10 shows individual Test 3 phrase scores plotted against those on Test 2 for the listeners receiving prosodic training (panel A) and phonemic training (panel B). Most of the higher Test 2 performers (> 25%) showed gains after training, while some of the poorer Test 2 performers (≤ 25%) showed only small gains or slight declines. These trends were observed in both groups.

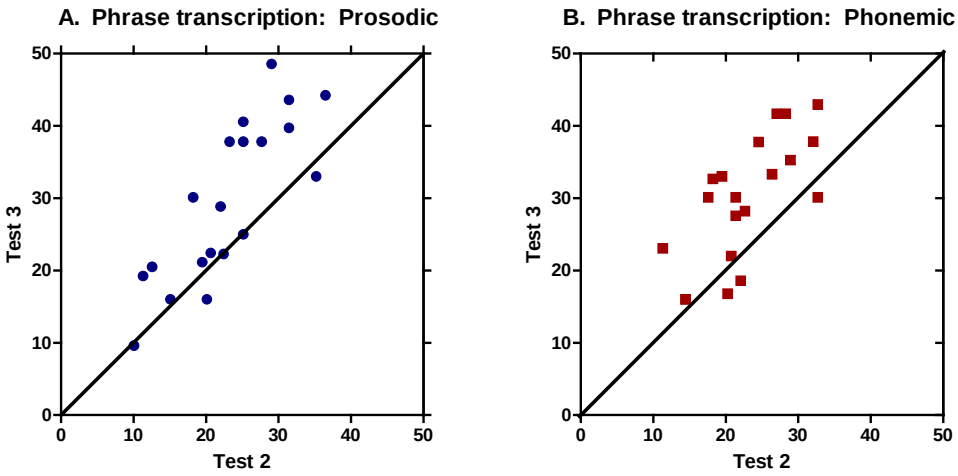



*Figure 10.* Scatterplots of individual phrase transcription scores on Test 2 vs. Test 3 for the listeners receiving degraded training with a) prosodic cues and b) phonemic cues.

Individual differences in the other segmental and suprasegmental measures of interest are shown in figure 11. Listeners receiving prosodic training are shown in the left column, and those receiving phonemic training are in the right column. Comparison of the IS/IW ratios in panels A and B reveals that several listeners in the prosodic group demonstrated large increases in their ratio of predicted to non-predicted insertion errors, while this was not the case for the listeners receiving phonemic training. Panels C and D show that more listeners demonstrated declines in consonant recognition in the prosodic group than the phonemic group. By contrast, panels E and F show that more listeners in the phonemic group saw declines in vowel recognition. Panels G and H show slightly more variability in sentence improvement for the listeners receiving prosodic training, compared to those receiving phonemic training.

Pearson correlation analyses were conducted to determine whether gains in phrase transcription were associated with gains in the other tasks. For the listeners receiving phonemic training, improvements in phrase accuracy were positively correlated with changes in IS/IW ratio ($r$ (20) = .551, $p$ = .012), and negatively correlated with vowel recognition ($r$ (20) = -.456, $p$ = .043). For the listeners receiving prosodic training, there were no significant correlations between gain on phrases and gain on the other measures of interest (all $p \geq .355$).[7]

[7] Given the significant group differences in performance on the prosodic vs. the phonemic *training* tasks, Pearson correlation analyses were also conducted to determine whether gains in phrase accuracy were associated with overall performance on training. There were no significant correlations ($p \geq .095$).

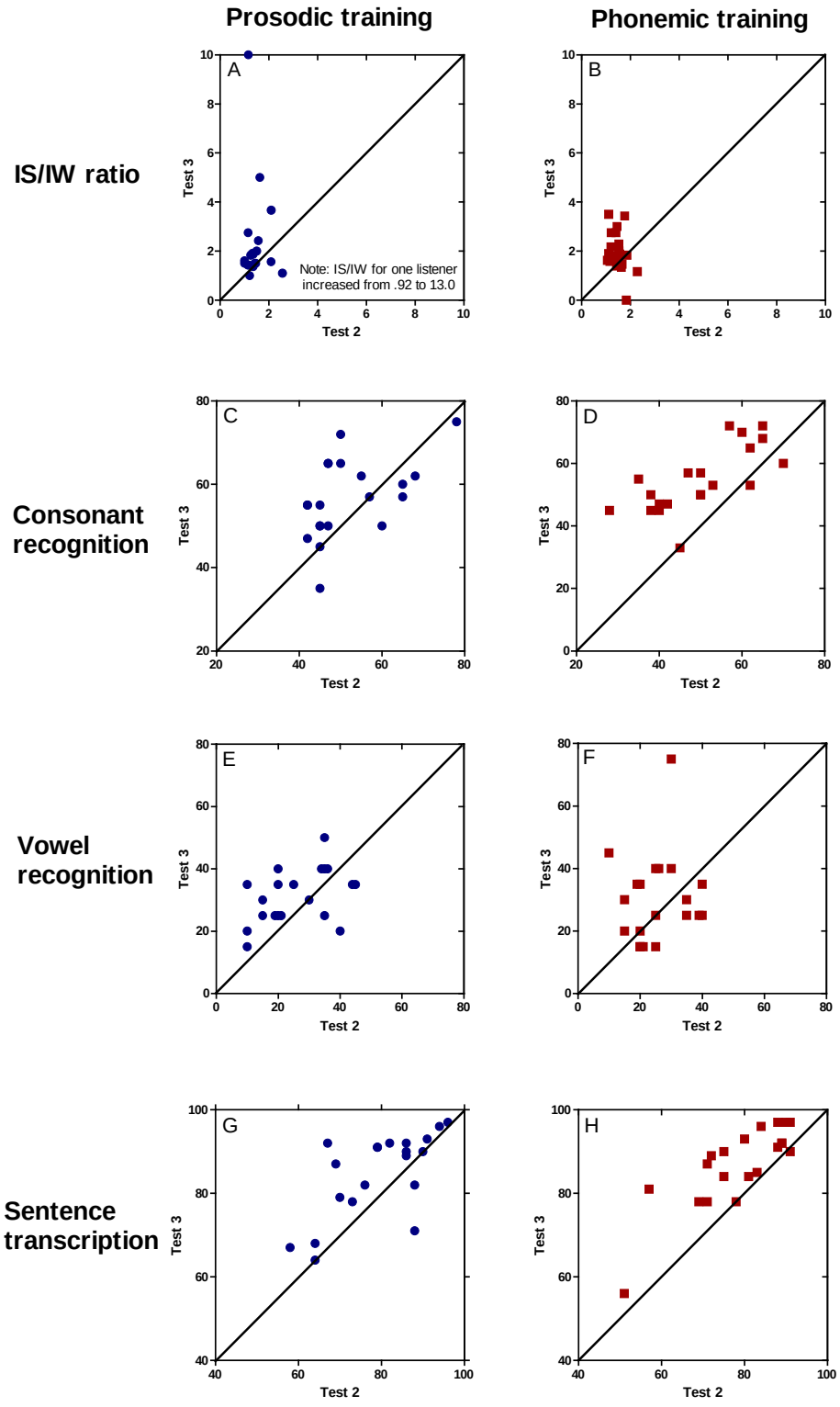



*Figure 11.* Scatterplots of Test 3 scores vs. those on Test 2. Not all measures are on the same scale; however, comparisons can be made between listening groups.

These trends motivated a closer examination of outcomes for individual listeners, given their training group assignment and individual differences in sensitivity to segmental, suprasegmental, and lexical cues. The test scores of 12 exemplar listeners were selected for qualitative comparison. Six were from the prosodic training group and six from the phonemic. In each group, three pairs of listeners representing the low, mid and high regions of Test 2 phrase accuracy were selected. At each of these points, one listener showed improved phrase accuracy after Training, and the other did not (figure 12).

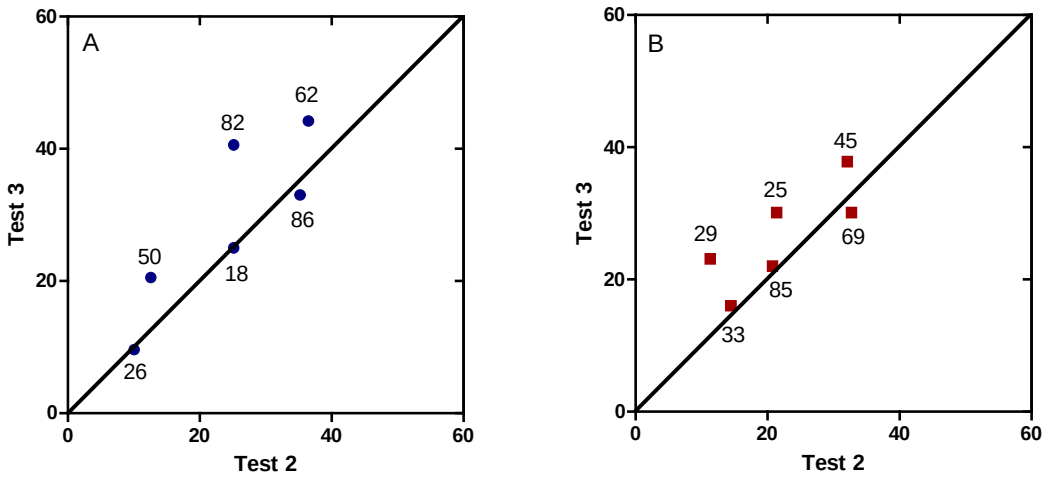



*Figure 12.* Test 3 vs. Test 2 phrase transcription scores for six exemplars from a) the prosodic training group, and b) the phonemic training group. Each listener is identified by a participant number.

Figure 13 illustrates differences in the clusters of speech skills across each pair of listeners. Those receiving prosodic training are shown in Panel A; those receiving phonemic training are shown in Panel B. Listeners are identified by participant number. The grayscale bars represent Test 2 performance for the measures of interest: IS/IW ratio is shown by the black bars (left Y-axis). Consonant, vowel, and sentence scores are shown respectively by the medium gray, light gray, and white bars (right Y-axis). The stacked blue and red bars represent the direction of change after Targeted Training with either prosodic or phonemic cues. Comparison across the low-, mid-, and high-accuracy listeners in both panels reveals that listeners' performance on the phrase transcription task in Test 2 was generally reflected in their Test 2 performance on vowels and sentences (light gray and white bars).

Examination of the post-training gains in both panels reveals some trends that may illuminate the effect of Training Strategy for individual listeners. Low- and mid-performing listeners who did *not* make improvements on phrase accuracy showed only small gains or declines in the skill area in which they did *not* receive training. For example, Panel A shows that consonant recognition (dark gray bars) minimally improved for participant 26 and decreased for participant 18, neither of whom made gains in phrase accuracy after *prosodic* training. Panel B shows that IS/IW ratio (black bars) marginally increased for participant 33 and precipitously declined for participant 85, neither of whom demonstrated gains in phrase accuracy after *phonemic* training.

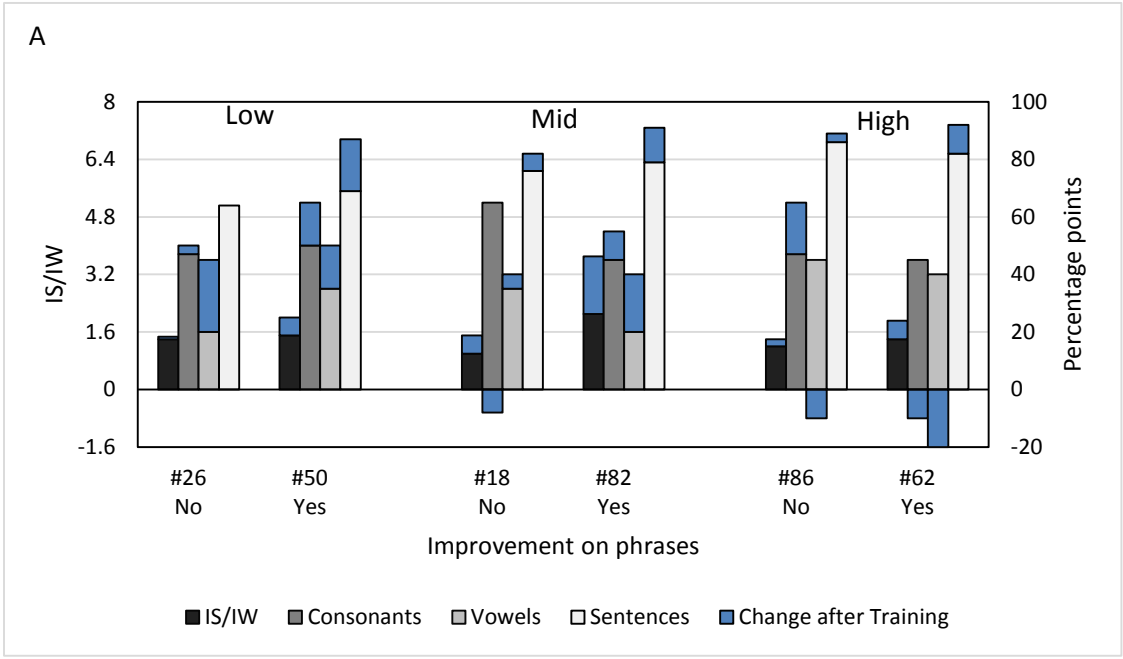
The low and mid listeners who *did* show gains in phrase accuracy after training demonstrated larger increases in several measures. For example, participants 50 and 82 in

the prosodic group showed gains in all four measures. In the phonemic group, participants 29 and 25 showed increases in IS/IW ratio and consonant recognition, as well as sentence transcription.

Differences across the higher performing listeners were not as clear-cut. In the prosodic group, participants 86 and 62 had similar pre-training profiles, with relatively high vowel and sentence scores. Participant 62 made improvements in phrase accuracy after training despite declines in both consonant and vowel recognition, and only a small increase in IS/IW ratio. Participant 86, on the other hand, made no improvements in phrase accuracy after training despite a relatively large gain in consonant recognition. In the phonemic group, the two higher-performing participants (69 and 45) also demonstrated similar pre-training profiles. Both had relatively high scores on consonants and sentences. After training, both showed gains in all four measures. While participant 45 demonstrated improvements in phrase accuracy, participant 69 did not.

*Summary and discussion of Targeted Training effects.* Group analyses showed specific training effects. On average, listeners demonstrated improvement on the skill area in which they received Targeted Training. Gains in phoneme discrimination were observed in the group receiving phonemic training with the degraded signal. By contrast, gains in prosodic discrimination were observed not only in the group receiving training with the degraded signal, but also in the control group who trained with the clear signal. Both groups receiving training with the degraded signal showed significant increases in use of the Metrical Segmentation Strategy. Listener profiles suggest that those whose phrase accuracy was higher also showed higher performance with sentences and

consonants, although this was not consistent across all listeners. Training did not consistently result in improved skills in targeted domain, as evidenced by small or absent changes in IS/IW ratio in some prosodic exemplars, and small improvements in consonant recognition in some phonemic exemplars.



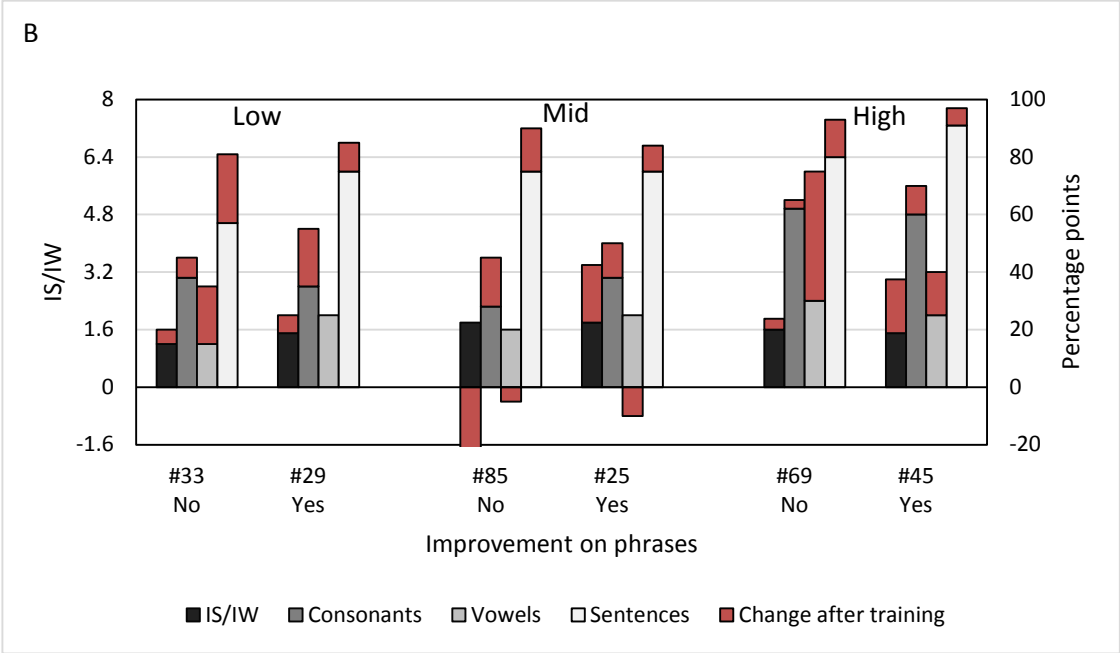




*Figure 13.* Speech skill profiles for exemplar listeners demonstrating low, mid, and high performance on the phrase transcription task in a) the prosodic training group and b) the phonemic training group.

## General Discussion

The present study investigated the effects of training specificity and individual differences on the perceptual learning of degraded speech. Before targeted training was assessed, the results of the non-guided Familiarization were examined. Results from the Familiarization task indicated that, on average, non-targeted practice facilitated learning, to a point. Sentence intelligibility plateaued toward the end of the Familiarization task. Post-Familiarization testing showed that mean accuracy on segmental tasks increased, and there was a corresponding decrease in overall use of suprasegmental cues to guide lexical segmentation. These findings align with those from other studies of the effects of non-guided familiarization on the perceptual learning of degraded speech (Borrie et al., 2012a, b).

There was, however, considerable variability in post-Familiarization outcomes. Lower- and higher-performing listeners showed contrasting effects of Familiarization, both in terms of overall phrase accuracy and use of syllable strength cues for lexical segmentation. An absence of clear associations between phrase accuracy and ability to use lower-tier segmental or suprasegmental cues led to a more qualitative approach to exploring differences. Examination of exemplar pairs representing different points along the performance continuum suggests that listeners with poorer performance on the phrase transcription task also had relatively low ability to use both segmental and lexical cues, even after Familiarization. Those demonstrating higher performance on phrases also had higher scores on other measures, although the specific areas of strength differed across listener. These findings resemble those reported by Choe et al. (2012) in their study of

individual differences in the perceptual strategies used by individuals transcribing dysarthric speech. They also suggest that the benefits of non-guided practice vary widely across individuals.

Given the gains made with non-guided practice with degraded speech, the Targeted Training task allowed for the examination of further specific training effects. Several questions were explored; findings related to each question are discussed below.

*Can targeted training increase acuity to degraded segmental or suprasegmental cues?* Analyses of training-specific effects indicated that acuity to the targeted cue did increase after training. The group receiving phonemic training with the degraded signal showed gains in their ability to discriminate phonemic cues, while the group receiving prosodic training showed gains in prosodic acuity. Interestingly, the listeners who heard the clear signal also showed improved ability to discriminate degraded prosodic contours after training. The lower overall performance on the prosodic training task, regardless of the type of signal heard during training, may speak to this. Given the strong focus on phonemic awareness experienced by many students in their elementary years, listeners in the present study may initially have had more difficulty shifting attention toward the prosodic features in the stimuli.

*Does increased acuity to segmental or suprasegmental cues generalize to improvements on other speech tasks?* Results of these analyses are equivocal. In the suprasegmental domain, the ability of listeners to generalize learning of word-level prosodic contours to the discrimination of sentence-level contours is unclear. While the post-training scores on the question/statement task were highest for the group receiving

degraded prosodic training, the difference among training groups did not reach significance. This is perhaps because of the high pre-training performance on this task. Possible ceiling effects might be avoided with a more complex task involving additional prosodic contrasts, beyond rising/falling. For example, tasks involving the location of a stressed word in a sentence or the identification of emotional prosody could offer greater insight into this question.

In the segmental domain, listeners receiving training with the degraded signal showed improved recognition of consonants, but not vowels. This outcome is not wholly unexpected. While the word triplets used in Targeted Training were designed to provide exposure to five different vowels, vowel identification per se was not targeted during training. In addition, the vowel test included five additional vowels not present in the training stimuli. On the other hand, all 20 items in the consonant recognition test appeared in the training stimuli. That both the degraded phonemic *and* prosodic groups showed improvements in consonant recognition suggests that improved phoneme discrimination was necessary for improved consonant identification.

*Does targeted training lead to the increased use of suprasegmental cues to parse degraded speech?* The groups who received training with the degraded signal (both prosodic and phonemic) demonstrated significant increases in their ratios of predicted to non-predicted lexical boundary insertion errors. These results indicate an increased use of the Metrical Segmentation Strategy to parse low-context phrases after training.

The use of syllable strength cues to guide lexical segmentation has been described as occurring when access to higher-level lexical and segmental information is insufficient

for lexical boundary identification (Choe et al., 2012). This implies an inverse relationship between segmental acuity and use of the MSS: listeners are driven toward a reliance on syllable strength cues when segmental units are more difficult to discern. In the present study, the groups trained with the degraded signal showed improved segmental (consonant) scores *and* increased use of the MSS after training. While phrase accuracy was not significantly higher for these groups, mean scores did show trends in that direction. These results suggest that targeted training may have increased listeners' flexibility in cue use; they were able to draw from both segmental and suprasegmental tiers for lexical segmentation.

*Is passive exposure sufficient to increase acuity to and use of these cues, or is active attention necessary?* Neither the phonemic nor the prosodic group demonstrated improved ability to discriminate the non-targeted cue after training. This suggests that passive exposure is not sufficient to improve the ability to resolve phonemic or prosodic differences, at least for the discrimination task employed in the present study. On the other hand, both the prosodic *and* the phonemic groups showed increased use of prosodic (suprasegmental) cues for lexical segmentation. This indicates that implicit learning of suprasegmental elements, at least for the purposes of word boundary identification, can occur even when attention is not explicitly directed to them. This is reminiscent of findings by Chandrasekaran et al. (2014) regarding Mandarin tone learning by naïve listeners. They observed that tone learning appeared to be an implicit, "reflexive" process that was most effectively supported by immediate feedback that was sparse, i.e., absent explicit rules for classifying tone categories. One question arising from these findings is whether the learning of other types of speech categories (such as consonants)

might be more explicit or "reflective." If this were the case, targeted training on phonemic differences using stimuli that also "bombard" the listener with prosodic variation could yield the largest overall benefit.

*How do individual differences in perceptual cue weighting affect training outcomes?* The large amount of individual variation in outcomes highlights the importance of looking beyond the means when investigating the effects of training. It also highlights the challenges involved in characterizing individual differences. Correlation analyses did not provide evidence of strong associations between specific sublexical skills and training outcomes. The strongest predictor of phrase accuracy appeared to be performance on the sentence task. This suggests that the ability to use lexical knowledge was tied particularly closely to overall performance in the low-context phrases. Beyond higher-level lexical knowledge, the different clusters of skills observed in the exemplar pairs indicates that individual listeners achieved similar levels of performance by leveraging different combinations of skills. The lack of improvement in phrase accuracy by some exemplars was accompanied by an absence of gain in the targeted skill area. This begs the question of whether those listeners would have shown greater post-training improvements had they been assigned to the other Training Strategy. In addition, the lack of improvement in sentence accuracy shown by the lowest performing exemplar suggests that training strategies targeting the use of higher level linguistic and contextual cues, rather than lower level segmental or suprasegmental cues, may provide more immediate benefit to some listeners.

A comparison of the exemplars who appeared in both the Familiarization and Training analyses indicates that some individuals benefitted from both non-guided Familiarization and Targeted Training, while others showed benefit from only one or the other. This is further evidence for the degree of individual variation in learning, and provides additional support for an individualized approach to intervention for cochlear implant users. A within- subjects design that parallels the current experiment could parlay into a test battery to predict patient outcomes.

*Clinical implications and future directions*

The wide variation in individual outcomes observed in the present study points to the potential importance of a personalized approach to aural rehabilitation. Listeners who already show higher performance with segmental tasks, for example, might benefit instead from training with suprasegmental cues. Those who show difficulty using contextual or lexical cues might receive more benefit from higher-level linguistic training, rather than focusing on sub-lexical cues.

Studies of the effectiveness of clinical intervention for individuals with hearing loss cite the importance of addressing skills beyond solely the auditory domain, including use of communication strategies (Tye-Murray, 1991, 1992), cognitive processing (Tao et al., 2014), and auditory-visual speech perception (Gagne and Wyllie, 1989). Efforts to profile listeners based on their communication strengths and needs should include broader-ranging processes such as these (Demorest and Erdman, 1987). While larger and more complex data sets likely increase the challenges associated with identifying

associations between variables, they also improve the likelihood of capturing meaningful trends. A large-scale data analytics approach, such as that used in teaching and learning research (U.S. Department of Education, 2012) could be useful in this regard. Given that the number of adult cochlear implant users in the United States now exceeds 40,000 (NIH Factsheet, 2010), large-scale studies are not outside the realm of possibility, assuming a shared spirit of collaboration by manufacturers, clinical sites, research groups, and of course cochlear implant users themselves.

References

ANSI (2004). ANSI S3.21 - 2004, Methods for manual pure-tone threshold audiometry. *American National Standards Institute*, New York.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., & Kishon-Rabin, L. (1988). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. *Ear and Hearing, 9,* 306-312.

Borrie, S.A., McAuliffe, M. J., Liss, J.M., Kirk, C., O'Beirne, G. A., & Anderson, T.J. (2012a). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes*, *27,* 1039-1055.

Borrie, S.A., McAuliffe, M. J., Liss, J.M., O'Beirne, G. A., & Anderson, T.J. (2012b). A follow-up investigation into the mechanisms that underlie improved recognition of dysarthric speech. *Journal of the Acoustical Society of America, 132,* 102-108.

Brown, C.A., & Bacon, S.P., (2009). Achieving electric–acoustic benefit with a modulated tone. *Ear and Hearing, 30,* 489–493.

Chandrasekaran, B., Sampath, P.D., & Wong, P.C.M. (2010). Individual variability in cue-weighting and lexical tone learning. *Journal of the Acoustical Society of America, 128,* 456-465.

Chandrasekaran, B., Yi, H.G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review, 21,* 488-495.

Choe, Y.-K., Liss, J.M., Azuma, & T., Mathy, P. (2012). Evidence of cue use and performance differences in deciphering dysarthric speech. *Journal of the Acoustical Society of America, 131,* 112-118.

Clopper, C.G., & Pisoni, D.B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47,* 207-239.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31,* 218-236.

Cutler, A., & Carter, D.M. (1987). The predominance of strong syllables in the English vocabulary. *Computer Speech and Language, 2,* 133-142.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology - Human Perception and Performance, 14,* 113-121.

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, 134,* 222-241.

Demorest, M., & Erdman, S., (1987). Development of the communication profile for the hearing impaired. *Journal of Speech and Hearing Disorders, 52,* 129-143.

Eady, S.J., & Cooper, W.E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America, 80,* 402-15.

Fear, B.D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America, 97,* 1893-1904.

Fu, Q.-J., Chinchilla, S., & Galvin, J. (2004). The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. *JARO – Journal of the Association for Research in Otolaryngology, 5,* 253-260.

Fu, Q.-J., & Galvin J.J. III (2006). Recognition of simulated telephone speech by cochlear implant users. *American Journal of Audiology, 15,* 127-132.

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

Gagne JP, & Wyllie KA. (1989). Relative effectiveness of three repair strategies on the visual identification of misperceived words. *Ear & Hearing, 10,* 368-74.

Gifford, R. H., Shallop, J.K., & Peterson, A.M. (2008). Speech recognition materials and ceiling effects, considerations for cochlear implant programs. *Audiology and Neuro-otology, 13,* 193-205.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585-612.

Hazan, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Perception and Psychophysics, 49,* 187-200.

Helms Tillery, K. Brown, C.A., & Bacon, S.P. (2012). Comparing the effects of reverberation and of noise on speech recognition in simulated electric-acoustic listening. *Journal of the Acoustical Society of America, 313,* 416-423.

Helms Tillery, K. Brown, C.A., & Bacon, S.P. (2009, July). *Spectral resolution and intelligibility of reverberant speech in simulated electric-acoustic listening.* Poster presented at the Conference on Implantable Auditory Prostheses, Lake Tahoe, CA.

Hervais-Adelman, A., Davis, M.H., Johnsrude, I., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology, 34,* 460-474.

Kessler, B. & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language, 37,* 295-311.

Krull, V., Luo, X., & Iler Kirk, K. (2012). Talker-identification training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition. *Journal of the Acoustical Society of America, 131,* 3069-3078.

Li, T., Galvin, J.J. III, & Fu, Q.-J. (2009). Interactions between unsupervised learning and the degree of spectral mismatch on short-term perceptual adaptation to spectrally shifted speech. *Ear and Hearing,* 30, 238-249.

Liss, J.M., Spitzer, S., Caviness, J.N., Adler, C.A., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *Journal of the Acoustical Society of America,* 104, 2457-2466.

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

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

Loebach, J.L., Pisoni, D.B., & Svirsky, J.A. (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 and Performance,* 36, 224-234.

Mattys, S.L., White, L., & Melhorn, J.F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology - General, 134,* 477-500.

Mattys, S.L., Davis, M. H., Bradlow, A.R., & Scott, S.K. (2012). Speech recognition in adverse conditions. *Language and Cognitive Processes, 27,* 953-978.

Meister, H., Landwehr, M., Pyschny, V, Wagner, P, & Walger, M. (2011). The perception of sentence stress in cochlear implant recipients. *Ear and Hearing, 32,* 459-467.

Murry, T., Brown, W.S. Jr., & Morris, R.J. (1995). Patterns of fundamental frequency for three types of voice samples. *Journal of Voice, 9,* 282-289

National Institutes of Health, National Institute on Deafness and Other Communication Disorders (2010). *Cochlear implants.* Retrieved 4/2/2015 from http://report.nih.gov/nihfactsheets/ViewFactSheet.aspx?csid=83

Nelson, P.B., Jin, S., Carney, A.E., & Nelson, D.A. (2003) Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America, 113,* 961-968.

Peng, S.-C., Lu, N., & Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiology and Neuro-otology, 14,* 327-337.

Peng, S.-C., Chatterjee, M., & Lu, N. (2012). Acoustic cue integration in speech intonation recognition with cochlear implants. *Trends in Amplification, 16,* 67-82.

Perrachione T.K., Lee, J., Ha, L.Y.Y., & Wong, P.C.M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America, 130,* 461-472.

Poissant, S. F., Whitmal, N. A., & Freyman, R. L. (2006). Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *Journal of the Acoustical Society of America, 119,* 1606-1615.

Rogers, C.F., Healy, E.W., & Montgomery, A.A. (2006). Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions. *Journal of the Acoustical Society of America, 119,* 2276-2287.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106,* 3629-3636.

Samuel, A.G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention Perception & Psychophysics, 71,* 1207-1218.

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 environmental sounds. *Trends in Amplification, 16,* 83-101.

Spitzer, S. M., Liss, J. M., & Mattys, S. L. (2007). Acoustic cues to lexical segmentation: A study of resynthesized speech. *Journal of the Acoustical Society of America, 122,* 3678–3687.

Spitzer, S., Liss, J., Spahr, T., Dorman, M., & Lansford, K. (2009) The use of fundamental frequency for lexical segmentation in listeners with cochlear implants *Journal of the Acoustical Society of America, 125*, 236-241.

Stacey, P.C. & Summerfield, A.Q. (2008). Comparison of word- sentence- and phoneme-based training strategies in improving the perception of spectrally distorted speech. *Journal of Speech, Language, and Hearing Research, 51,* 526-538.

Surprenant A.M., & Watson, C.S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *Journal of the Acoustical Society of America, 110,* 2085-2095.

Tao D., Deng, R., Jiang, Y., Galvin, J.J. III, & Fu Q-J, (2014). Contribution of auditory working memory to speech understanding in Mandarin-speaking cochlear implant users. *PLoS ONE 9,* e99096.

Tye-Murray, N. (1991). Repair strategy usage by hearing-impaired adults and changes following communication therapy. *Journal of Speech-Language-Hearing Research, 34*, 921-928.

Tye-Murray, N. (1992). Preparing for communication interactions: the value of anticipatory strategies for adults with hearing impairment. *Journal of Speech-Language-Hearing Research, 35,* 430-435.

U.S. Department of Education, Office of Educational Technology (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief*, Washington, D.C.

Watson, C.S. (1980). Time course of auditory perceptual learning. *Annals of Otology Rhinology and Laryngology, 89 (Suppl 74),* 96-102.

Watson, C.S. (1991). Auditory perceptual learning and the cochlear implant. *American Journal of Otology, 12,* 73-79.

Wayne, R., & Johnsrude, I. (2012, February). *Is perceptual learning of noise-vocoded speech enhanced by audiovisual speech information?* Poster presented at the 35th Annual Midwinter Meeting of the Association for Research in Otolaryngology, San Diego, CA.

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: a remarkable past and a brilliant future. *Hearing Research, 242,* 3-21.

Zhang, T., Dorman, M.F., Fu, Q.-F., & Spahr, A.J. (2012). Auditory training in patients with unilateral cochlear implant and contralateral acoustic stimulation. *Ear and Hearing, 33,* e70-e79.

# APPENDIX A

## COMPUTERIZED AUDITORY TRAINING PROGRAMS FOR ADULT CI USERS

| Product | Company | Type of Training Stimuli |
|---|---|---|
| Seeing and Hearing Speech | Sensimetrics | Vowels, consonants**, stress, intonation, length,** and everyday communication |
| eARena | Siemens | Adaptive training at word and sentence level; **pitch, loudness and duration discrimination**. |
| Angel Sound (formerly Sound and Way Beyond) | Cochlear Americas | Interactive computer training activities with vowel, consonant, sentence, telephone, and music stimuli |
| SoundScape | Med-El | Sentence level with options to adjust amount of noise, rate of speech, or gender of speaker. Telephone training activity also available |
| Speech Perception Assessment and Training System (SPATS) | Communication Disorders Technology | Predominantly syllable training using 100 of the most important sounds for speech perception; sentences from multiple talkers. Training in quiet or noise available. |
| The Listening Room (CLIX) | Advanced Bionics | Interactive listening at word and sentence level with discrimination and identification activities; telephone and music training options available. |
| Listening and Communication Enhancement | Neurotone | Adaptive training at sentence level with background noise, competing talkers, fast speech; auditory memory and auditory cloze activities. |

# APPENDIX B

# ASU INSTITUTIONAL REVIEW BOARD FOR HUMAN SUBJECTS RESEARCH

# APPROVAL DOCUMENTATION

ASU Knowledge Enterprise Development

APPROVAL:CONTINUATION

Julie Liss
Health Solutions, College of
480/965-9136
JULIE.LISS@asu.edu

Dear Julie Liss:

On 4/17/2014 the ASU IRB reviewed the following protocol:

| | |
|---|---|
| Type of Review: | Modification and Continuing Review |
| Title: | Perceptual Learning of Spectrotemporally Reduced Speech |
| Investigator: | Julie Liss |
| IRB ID: | 1305009211 |
| Category of review: | (4) Noninvasive procedures, (7)(b) Social science methods, (7)(a) Behavioral research |
| Funding: | Name: NIDCD: National Institute on Deafness and other Communication Disorders ; Funding Source ID: HHS-NIH-NIDCD-National Institute of Deafness & Other Communi, |
| Grant Title: | None |
| Grant ID: | None |
| Documents Reviewed: | None |

The IRB approved the protocol from 4/17/2014 to 5/15/2015 inclusive. Three weeks before 5/15/2015 you are to submit a completed "FORM: Continuing Review (HRP-212)" and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 5/15/2015 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

# APPENDIX C

## WORD TRIPLET STIMULI

The word triplet corpus contained 260 tokens; 200 were used for targeted training, and 20 for each test. Each triplet contained a phonemic and a prosodic contrast (e.g., Bill. Fill? Bill?). This structure exposed listeners to both contrasts, while allowing targeted practice (or testing) with just one of them.

The phonemic contrast was always word-initial, and comprised differences in place, manner, and/or voicing. The word-initial phonemes were: /p, b, m, w, f, v, voiced th, t, d, n, s, z, l, r, sh, ch, dj, k, g, h/. Across the corpus, the frequency distribution of these phonemes roughly resembled that of word-initial consonants in English CVCs (Kessler and Treiman, 1997). One word in each triplet contained a prosodic contour that contrasted with the prosody in the other two words. Contours were either rising or falling. Intensity and duration co-varied naturally with F0 variation.

Words were audio recorded individually using Audacity (sourceforge.net). To ensure that the prosodic contours were clearly discernible, two independent judges rated each production on a 7-point scale (1=falling, 4=unsure, 7=rising). Productions receiving a rating of 3, 4, or 5 were re-recorded. Once all contours were validated, individual audio files were combined to form triplets using Audacity. Items within a triplet were separated by 300 ms of silence. The phonemic and prosodic contrasts fell approximately equally in each position (e.g., AAB, ABA, BAA), such that, across the entire corpus, both contrasts fell on the same word one third of the time.

The corpus contained five sets of triplets, each with a different word ending (-ill, -ooze, -an, -od, -um). The vowels represented different positions in English vowel space (high front, high back, low front, low back, and mid). The final consonants represented various manners of production (liquid, fricative, stop, nasal), and were all voiced. Within each sub-set, half of the triplets were produced by the male talker and half by the female. For training, the stimuli were blocked by talker, then by word ending. Talker and set order were randomized across listener, and the order of the triplets within each set was also randomized. During testing, the talker, word ending, and triplet order were randomized.

| -ill | -ooze | -an | -od | -um |
|---|---|---|---|---|
| bill | booze | ban | bod | bum |
| chill | coos | can | cod | come |
| dill | choose | Dan | god | chum |
| gill | whose | fan | mod | dumb |
| hill | lose | Jan | nod | gum |
| Jill | rues | man | pod | hum |
| mill | sues | Nan | rod | mum |
| nil | shoes | pan | sod | numb |
| pill | twos | ran | shod | rum |
| sill | woos | tan | Todd | sum |
| till | zoos | than | wad | tum |
| will | | van | | |

# APPENDIX D

## QUESTION/STATEMENT STIMULI

The question/statement corpus contained 60 five-word sentences, used during the testing sessions. Each was produced with falling or rising prosody to indicate a statement or a question. To ensure that the prosodic contours were clearly discernible, two independent judges rated each production on a 7-point scale (1=statement, 4=unsure, 7=question). Any productions receiving a rating of 3, 4, or 5 were re-recorded and re-validated.

He baked a small cake
Your bike tire was slashed
Those new boots were stiff
They bought a used truck
I broke the wine glass
Their mom built that house
His cat caught a bird
He caught you off guard
She checked out a book
Your dad will work late
You fed the big fish
The golf club was bent
My green pants are clean
You heard the good news
The hen laid an egg
That boy jumped the ditch
The kids want to swim
She won't paint her nails
They like to pick grapes
The rose was bright pink

The sale will end soon
Those dress shoes are blue
The soup still needs salt
Those steel bars won't rust
The storm was not mild
You'll take a long trip
His voice was quite hoarse
They watched the old film
That white shirt is stained
She wrote him a note
The bath mat is wet
The ground beef looked brown
His row boat has cracks
The school bus is slow
The trash can was gone
Her race car went fast
The sly cat will leap
The grilled cheese is hot
My tea cup did break
The lost dog found home

He gave a good speech
The house was on fire
His ice cream might melt
A lunch can have germs
The nurse is on call
The jet plane took off
This short plant could grow
His tea pot will boil
She put you on hold
My ripped shorts are fixed
The ski slope is steep
A bee sting will hurt
The iced tea is warm
The old tire was flat
The train was on time
The ash tree grew leaves
The dump truck is full
That wash tub could leak
They walked to the mall
The grey wolves are fierce

# APPENDIX E

## EXPERIMENTAL TASK INSTRUCTIONS TO PARTICIPANTS

*Phoneme discrimination*
You will hear 3 words. Each word will be represented by a number on the screen. One of the words will start with a consonant sound that is different from the other two words. Type the number of the word that starts with a different consonant sound, and press Enter.

*Prosodic contour discrimination*
You will hear 3 words produced with rising or falling voice pitch contours. Each word will be represented by a number on the screen. One of the words will have a different pitch contour than the other two words. Type the number of the word with the different pitch contour, and press Enter.

*Question/statement discrimination*
You will hear some sentences. In each one, the talker will ask a question or make a statement. After each sentence is presented, type 'q' if it was a question, or 's' if it was a statement. Then press Enter to hear the next sentence.

*Consonant recognition*
You will hear some nonsense words, listed below. After each word is presented, find the number of the word you heard, and type it using the keypad. Then press Enter to hear the next word.

1. aba
2. acha
3. ada
4. afa
5. aga
6. aha
7. aja
8. aka
9. ala
10. ama
11. ana
12. apa
13. ara
14. asa
15. asha
16. ata
17. atha
18. ava
19. awa
20. aza

*Vowel recognition*
You will hear some words, listed below. After each word is presented, find the number of the word you heard, and type it using the keypad. Then press Enter to hear the next word.

1. had
2. hawed
3. hayed
4. head
5. heed
6. hid
7. hoed
8. hood
9. hud
10. who'd

*Sentence transcription*
You will hear some sentences. After each sentence is presented, type all the words you heard. It's okay to guess. If you don't have any idea what a word is, type an X for that word. Correct any typos, then press Enter to hear the next sentence.

*Phrase transcription*
You will hear some phrases. Each phrase contains real English words, although the phrase may not make sense. After each phrase is presented, type all the words you heard. It's okay to guess. If you don't have any idea what a word is, type an X for that word. Correct any typos, then press Enter to hear the next phrase.

*Familiarization*
You will hear some sentences. After each sentence is presented, type all the words you heard. It's okay to guess. If you don't have any idea what a word is, type an X for that word. Correct any typos, then press Enter to hear the sentence again and read the answer.

*Targeted training (Phonemic)*
You will hear 3 words. Each word will be represented by a number on the screen. One of the words will start with a consonant sound that is different from the other two words. Type the number of the word that starts with a different consonant sound, and press Enter. After you press Enter, you will hear the words again, and the correct answer will be highlighted in green on the screen.

*Targeted training (Prosodic)*
You will hear 3 words produced with rising or falling voice pitch contours. Each word will be represented by a number on the screen. One of the words will have a different pitch contour than the other two words. Type the number of the word with the different pitch contour, and press Enter. After you press Enter, you will hear the words again, and the correct answer will be highlighted in green on the screen. Rising pitch will be shown with a question mark (?) and falling pitch will be shown with a period (.).