11-1-2017

Generalization of perceptual learning of degraded speech across talkers

Julia Jones Huyck
Queen's University, Centre for Neuroscience Studies, Kingston

Rachel H. Smith
University of Cambridge

Sarah Hawkins
University of Cambridge

Ingrid S. Johnsrude
Queen's University, Centre for Neuroscience Studies, Kingston, ijohnsru@uwo.ca

Generalization of perceptual learning of degraded speech across talkers

Julia Jones Huyck[a]

Department of Psychology and Centre for Neuroscience Studies, Queen's University,

62 Arch Street, Kingston Ontario, Canada K7L 3N6

and

Speech Pathology and Audiology Program, Kent State University[b]

1325 Theatre Drive, Kent, OH, USA 44242

Rachel H. Smith

Department of Linguistics, University of Cambridge,

Cambridge, United Kingdom

and

Glasgow University Laboratory of Phonetics, University of Glasgow [b],

Glasgow, United Kingdom

Sarah Hawkins

Department of Linguistics and Centre for Music and Science [b], University of Cambridge,

Cambridge, United Kingdom

Ingrid S. Johnsrude

Department of Psychology and Centre for Neuroscience Studies, Queen's University,

Kingston Ontario, Canada

The Brain and Mind Institute, University of Western Ontario [b],

London, Ontario, Canada

[a]Corresponding author: jhuyck@kent.edu

[b]Current affiliation

Abstract

Purpose: We investigated whether perceptual learning of noise-vocoded (NV) speech is specific to a particular talker or accent.

Method: Four groups of listeners (n=18 per group) were first 'trained' by listening to 20 NV sentences that had been recorded either by a talker with the same native accent as the listeners or a different regional accent. They then heard 20 novel NV sentences from either the native- or non-native-accented talker (test), in a 2x2 (training talker/accent x test talker/accent) design.

Results: Word-report scores at test for participants trained and tested with the same (native- or non-native-accented) talker did not differ from those for participants trained with one talker/accent and tested on another.

Conclusions: Learning of NV speech generalized completely between talkers. Two additional experiments confirmed this result. Thus, when listeners are trained to understand NV speech, they are not learning talker- or accent-specific features but instead are learning how to use the information available in the degraded signal. The results suggest that people with cochlear implants, who experience spectrally degraded speech, may not be too disadvantaged if they learn to understand speech through their implant by listening primarily to just one other talker, such as a spouse.

I. Introduction

People are remarkably good at comprehending speech even though the acoustic realization of a given utterance can vary markedly. This variability can arise from environmental factors (e.g., background noise and reverberation), from attributes of the medium used for communication (e.g., reduced frequency bandwidth over the telephone), from talker characteristics (e.g., age, sex, size, and regional accent), and from situational factors (e.g., pragmatic context and emotional state). The ability of most listeners to understand speech in its many acoustic realizations arises in part from perceptual learning; from experience-related improvements in the comprehension of unusual-sounding, accented, or degraded speech.

The intelligibility of degraded speech improves within the first few minutes of experience (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011). Artificial degradations, such as noise-vocoding, have been particularly useful for examining perceptual learning, since exposure can be precisely controlled. Another reason to study noise vocoding specifically is that the algorithm, which removes most of the fine spectral information while leaving the temporal structure largely intact, is similar to that implemented in cochlear implants (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). A better understanding of what, exactly, is learned as comprehension of noise-vocoded speech improves may therefore inform cochlear-implant rehabilitation programs.

Perceptual learning is usually measured by testing generalization to novel materials. Learning to understand noise-vocoded (NV) speech from a particular talker appears to generalize completely to untrained (i.e., novel) sentences (Davis et al., 2005) and words (Hervais-Adelman

et al., 2008) spoken by that talker, and to an untrained frequency region (i.e., from vocoded speech generated using low-pass noise to vocoded speech generated using high-pass noise), at least when the information in each vocoded frequency band is consistent with the information in that frequency band in the original (clear) signal (i.e., when the vocoded speech has not been spectrally shifted)(Fu & Galvin, 2003). Learning also appears to generalize to untrained carrier signals (e.g., from noise-vocoded to pulse train-vocoded speech), but this generalization is only partial: that is, when trained with one carrier signal and tested with a different one, test performance was not as good as if listeners were tested using the same carrier with which they had been trained (Hervais-Adelman et al., 2011).

An interesting set of experiments by Dahan and Mead (2010) highlights the phonetic conditions under which generalization of learning of NV single words is observed. These authors demonstrated incomplete generalization when the syllable positions of phonemes within words were changed between training and testing (i.e., onset vs. coda), suggesting that listeners learn more about the pronunciation of particular allophones than about abstract phonemes when they learn to understand noise-vocoded speech. They also systematically manipulated the talker heard at training and test, reasoning that if allophonic variation matters, so might talker-specific variation. However, they found rather weak and inconsistent effects of talker: In one experiment, comprehension during testing was slightly better for words spoken by the same talker as was used for training, than for those spoken by an new talker. However, in a subsequent experiment, there was no significant difference between the comprehension of words from the trained and untrained talkers. Dahan and Mead proposed that this outcome might relate to the discriminability of the two voices, with greater generalization between easily discriminable voices (e.g., between male and female voices).

Here we systematically investigate the extent to which listeners who learn to understand noise-vocoded (NV) sentences with one talker generalize their learning to a different talker of the same sex. Unlike Dahan and Mead's (2010) study of generalization of learning of NV speech, we used sentences as our stimuli because most naturally occurring language involves utterances that express more-or-less complete thoughts, as opposed to single words. In Experiments 1 and 2, we examined whether learning to understand noise-vocoded (NV) speech is specific to the regional accent (and specific talker) used in training, or generalizes across accents (and talkers).

II. Experiment 1

Canadian listeners participated in one of three conditions: In the experimental "Switch" condition, two groups of listeners heard NV speech from two different talkers, one during training and the other during testing: they heard a young female talker with either a "native" Canadian English (CE) or "non-native" Standard Southern British English (SSBE) accent during training and sentences from the other talker (SSBE or CE accent) during testing. Naturally, each talker spoke in her own native accent; and hence "native"/"non-native" are used here with respect to the listeners' point of view. In one control condition (Constant condition), two groups were trained and tested with NV speech from the same talker (with either a native or non-native accent), thus maximizing the potential for learning to generalize to novel utterances at test. This condition served as a baseline against which to measure generalization to an untrained accent (in the Switch condition). In another control condition (Naive condition), listeners did not participate in the training session. They only heard the test NV sentences from either the native-accent or non-native-accent talker, in order to provide an estimate of baseline performance without previous training. Performance was measured as the proportion of words of each sentence

reported correctly. We compared percent-correct word-report scores among the three conditions to assess the extent to which learning of NV speech generalized between the two talkers (and accents).

If learning generalizes completely to a novel talker (and an untrained accent), then participants in the Switch condition should be indistinguishable from those in the Constant condition during test, and both groups should be significantly better than the Naive group, who receive no training. This complete generalization could be interpreted to mean that listeners can generalize what they learn about NV speech to novel talkers, even when those talkers have different native accents. If generalization is incomplete, then a robust difference between Constant and Naive groups should be observed during test (assuming the same-voice training is effective), but the difference between Switch and Naive groups should be smaller. This incomplete generalization could be interpreted to mean that perceptual learning of NV speech is specific to some, but not all, of the acoustical features that differ between talkers. Because only one talker per accent was used, it would be impossible to determine whether these features were accent-related, or related to other individual differences in speech acoustics.

A. Methods

We tested seventy-two students from Queen's University in Canada (54 females). All participants were between 18 and 25 years of age (mean = 19 years old, SD = 1.37 years). Participants were recruited through posters, email advertisement, and the Queen's Psychology 100 Subject Pool. The participants reported that Canadian English was their native accent and language (7 participants indicated that they had two native languages and 9 additional participants were fluent in at least one language other than English). They had normal self-

reported hearing, normal or corrected-to-normal vision, and no known history of language impairment. This study was cleared by the Queen's University General Research Ethics Board and written informed consent was received from all participants.

The experiment was organized into training and testing phases, presented sequentially with no break between them. During both phases, spectrally degraded (noise-vocoded; NV) sentences were presented one at a time. Sentences were each played once in distorted form, then in clear form, then in the distorted form again, since previous work demonstrates that such presentation seems to result in efficient learning (Davis et al., 2005). Participants were instructed to listen carefully to each sentence, and to write down as many words as they could understand immediately after they heard the first distorted presentation.

The stimuli used during training and testing consisted of 40 meaningful English sentences (e.g., "The whole sky was full of birds"). Each sentence was recorded by two female talkers in their early twenties: one was a native speaker of Canadian (Ontario) English and the other was a native speaker of SSBE. The forty sentences were split into four sets of ten each, matched for sentence duration (mean = 2028 ms, SD across sets = 97 ms) and the number of words per sentence (mean = 8.78 words, no variation across sets) as well as for naturalness and imageability (rated on a 7-point Likert scale by two groups of 18 participants; see Rodd, Davis, & Johnsrude, 2005). Two sentence sets (A & B) were used during training (with the order of the sets counter-balanced across participants) and two sets (C & D) were used during testing (also with counter-balancing across subjects). The cross-accent design minimized material effects because, across groups, each stimulus was used in all three (Constant, Switch, and Naïve) conditions.

Digital recordings were made in sound-attenuating booths and were subsequently downsampled to 22 kHz using Adobe CoolEdit software. Each of the forty sentences was noise vocoded using the method described by Shannon et al. (1995). First, each sentence was divided into six frequency bands selected to be approximately equally spaced along the basilar membrane (cut offs: 50, 229, 558, 1161, 2265, 4290 and 8000 Hz; Greenwood, 1990). Next, a smoothed amplitude envelope was extracted for each band. These envelopes were then used to modulate band-limited noises with the same cut-off frequencies (Shannon et al., 1995). Finally, the amplitude-modulated noises were recombined to form a new, degraded, sentence.

Listeners were split into six groups (3 conditions x 2 accents during the testing; n=12 per group). Two "Switch" groups were trained with NV speech either from the talker with a "non-native" (i.e., different from the accent of the listeners; SSBE) or a native (CE) accent and tested with different sentences from the talker with the other accent (native or non-native, respectively). Two "Constant" groups were trained and tested with only native-accented or non-native-accented NV speech, with only the sentences changing between training and testing. Because the voice and accent remained constant, performance of these groups provided an estimate of the maximum post-training performance that could be expected of any trained group for that accent; i.e., if learning generalizes completely from training to test in the Switch condition, performance levels should be similar to that in the Constant condition. Finally, in order to estimate baseline performance on the test NV sentences, two "Naïve" groups heard only the test sentences, in only a single accent (either native or non-native), and with no prior training sessions. If learning does not generalize at all from training to test in the Switch condition, performance level in this condition should be similar to that in the Naïve condition.

Before the experiment began, listeners were familiarized with the task and were screened to ensure that short-term memory capacity would not limit performance on the degraded sentences during the experiment. This was accomplished by giving all listeners four clear (un-degraded) sentences in their native accent, and asking them to perform the word-report task on these. To give participants an idea of the form of distortion presented in the experiment, they then listened to a highly intelligible native-accent NV sentence vocoded using 15 frequency bands and were asked to perform the word-report task on this.

Participants' responses were scored for the percentage of words in each sentence that were reported correctly. Words were considered correct if they matched the word in the sentence exactly and were reported in the correct order, even if intervening words were incorrect (Davis et al., 2005; Hervais-Adelman et al., 2008, 2011).

Training and testing data were analyzed separately and could not be compared to one another due to the different sentence sets used. If they had been compared, it would have been impossible to determine which effects were the result of item effects and which were due to the experimental manipulation. Word-report scores during training were analyzed using a mixed-design ANOVA with Time (2 levels: training trials 1-10 and training trials 11-20) as a within-subjects factor, Training Accent (2 levels: Native (CE) or Non-native (SSBE)) as a between-subjects factor, and Set order (2 levels: AB or BA) as a dummy variable. Word-report scores during testing were analyzed using a mixed-design ANOVA with Time (2 levels: testing trials 1-10 and testing trials 11-20) as a within-subjects factor, Condition (3 levels: Naive, Constant, or Switch) and Test Accent (2 levels: Native (CE) or Non-native (SSBE)) as between-subjects factors, and Set order (2 levels: CD or DC) as a dummy variable. We report the results of the statistical analyses performed on raw data ($\alpha = 0.05$) with Sidak adjustments; however, the

statistical conclusions did not change when the data were transformed into rationalized arcsine units (RAU; Studebaker, 1985) prior to performing the statistical tests.

B. Results

Learning during training was confirmed by a significant main effect of Time during Training ($F_{1,40} = 74.229$, $p < 0.001$). The groups trained with both the native (CE; Constant-Native and Switch-Native to Non-native Groups) and the non-native (SSBE; Constant-Non-Native and Switch-Non-native to Native Groups) voices improved their performance from the first 10 trials to the second 10 trials (post hoc testing, Native: $p < 0.001$, Non-native: $p < 0.001$). Such rapid learning over 10 sentences is entirely consistent with previous studies (Davis et al., 2005; Wayne & Johnsrude, 2012). During training (Figure 1), there was a near-significant Time x Condition x Training Accent interaction ($F_{1,2,40} = 4.052$, $p = 0.051$). Listeners who heard Native-accented NV speech outperformed those who heard Non-Native accented NV speech (main effect of Training Accent, $F_{1,40} = 20.118$, $p < 0.01$), indicating that features specific to the accent or voice were present in the degraded signal. There was a near significant interaction between Training Accent and Time ($F_{1,40} = 3.945$, $p = 0.054$), suggesting that listeners who heard Non-Native accented NV speech might have improved slightly more between Training trials 1-10 and Training trials 11-20 (mean change =17.629) than the listeners who heard Native-accented NV speech (mean change = 11.028). This may reflect the tendency for listeners who start worse to improve more, either due to regression towards the mean or ceiling effects.

During the test phase (Figure 2), none of the interactions involving Test Accent (including the three-way Condition by Test Accent by Time interaction) were significant ($p \geq 0.585$). However, there was a significant interaction between Condition and Time ($F_{1,40} = 6.510$,

$p < 0.001$). During the first ten test trials, word-report scores differed among listeners in the three conditions (post-hoc simple effect of Condition*:* $p < 0.001$). Specifically, the listeners in the Switch and Constant conditions demonstrated better word-report performance than Naïve listeners (both ps ≤ 0.028) and similar performance to one another ($p = 0.686$) at this first time-point. In the second half of testing (trials 11-20), the Naïve group was no longer different from the other two groups (post-hoc simple effect of Condition: $p = 0.846$), reflecting rapid learning in this group over the test trials.

In addition to the Condition by Time interaction, there were significant main effects of Time (performance improved between test trials 1-10 and test trials 11-20; $F_{1,68} = 23.279$, $p < 0.001$) and of Test Accent, with the listeners performing better with their native (CE) accent than with the non-native (SSBE) one ($F_{1,68} = 27.063$, $p < 0.001$). There was also a main effect of Condition ($F_{2,68} = 3.212$, $p = 0.047$), which was due to the fact that, as expected, participants in the Naïve condition had the worst performance and participants in the Constant condition performed the best. Posthoc pairwise comparisons revealed a trend towards a difference between the Constant and Naïve conditions ($p = 0.055$), with no other apparent differences (The main effect of Condition was obscured by the Condition x Time interaction discussed above).

To further investigate the Condition by Time interaction, three follow-up 2 Condition by 2 Time ANOVAs were conducted. They revealed that there was a significant Condition by Time interaction when the Constant condition was compared to the Naïve condition ($F_{2,68} = 17.042$, $p < 0.001$) and when the Switch condition was compared to the Naïve condition ($F_{2,68} = 11.239$, $p = 0.001$) but not when the Constant and Switch conditions were compared to one another ($F = 1.038$, $p = 0.312$).

Taken together, the results indicate that the Switch group, who heard training stimuli in one accent/voice followed by test stimuli in a different accent/voice, generalized at least as well as the Constant group, who heard the same talker throughout. Moreover, performance in both of these trained groups was superior to that in the Naïve group, reflecting their learning during training.

II. Experiment 2: Replication

To confirm the results of Experiment 1, we repeated part of the design in three groups of British participants. The methods and stimuli were the same as in Experiment 1, except that all listeners heard NV speech in their native (SSBE) accent during the post-training test.

A. Methods.

We tested thirty-six students from Cambridge University in the United Kingdom (21 females) who were between 18 and 26 years of age (mean = 21 years old, SD = 1.55 years). Participants were recruited through posters and email advertisement. All reported that Standard Southern British English (SSBE) was their native accent and language (one was also fluent in another language). They had normal self-reported hearing, normal or corrected-to-normal vision, and no known history of language impairment. The study was approved by the Humanities and Social Sciences Research Ethics Committee of the University of Cambridge, and all participants provided written informed consent.

Methods and stimuli were the same as described above, except that listeners were only tested with NV speech in their native accent. The British listeners were therefore split into only three groups (3 conditions x 1 accent during the testing; n=12 per group). The “Switch” group

was trained with NV sentences from the talker with a "non-native" (i.e., different from the accent of the listeners; CE) accent and tested using NV sentences from the talker with the "native" (SSBE) accent. The "Constant" group was trained and tested with only native-accented sentences. The "Naïve" group received no training and was tested only with the native accent.

B. Results

Learning during training (Figure 3) was confirmed by a significant main effect of time ($F_{1,20} = 20.22$, $p < 0.001$), reflecting significant improvement between the first and last 10 trials of training in the Constant Group (trained with native (SSBE) voice; $p = 0.002$) and the Switch Group (trained with non-native (CE) voice; $p = 0.010$). Unlike in Experiment 1, performance between listeners who heard the native vs. non-native voices did not differ (Accent x Time interaction: $F_{1,20} = 0.20$, $p = 0.657$; main effect of Accent: $F_{1,20} = 3.29$, $p = 0.085$). If anything, there was a small trend for performance to be better with the Non-native (CE) voice than with the Native (SSBE) one.

Since the purpose of this experiment was to see if the main results from the Canadian listeners could be replicated, all further statistical tests were one-tailed. As in the test phase of Experiment 1, there was a significant interaction between Condition and Time ($F_{2,30} = 3.024$, $p = 0.032$; Figure 4) in the test phase of Experiment 2. During the first ten testing trials, performance differed among listeners in the three conditions (simple effect of Condition*:* $F_{2,30} = 3.562$, $p = 0.021$). At this first time-point, the listeners in the Switch condition demonstrated better word-report performance than Naïve listeners ($p = 0.032$) and performance in the Constant condition was marginally greater than in the Naïve condition ($p = 0.055$); the Constant and Switch groups did not differ from one another ($p = 0.50$). This result is consistent with the results of Experiment 1: There was generalization from training to test in the Switch group, and both of the

trained groups (Switch and Constant) appeared to learn during training. As in Experiment 1, the Naïve group ceased to be different from any other group during the second half of testing (trials 11-20), indicating that this group learned rapidly over the test trials (post-hoc simple effect of group and all paired comparisons: all $p \geq 0.394$). Overall, there was improvement for all groups between test trials 1-10 and test trials 11-20 (main effect of Time; $F_{1,30} = 13.115$, $p < 0.001$). The main effect of Condition approached but did not reach significance ($F_{1,30} = 2.384$, $p = 0.055$), likely due to the smaller number of participants compared to the original experiment.

As in Experiment 1, to further explore the Condition by Time interaction, three 2 Condition x 2 Time ANOVAs were conducted. As before, there was a significant Condition by Time interaction when the Constant condition was compared to the Naïve condition ($F_{1,40} = 5.055$, $p = 0.018$) and when the Switch condition was compared to the Naïve condition ($F_{1,40} = 3.822$, $p = 0.033$) but not when the Constant and Switch conditions were compared to one another ($F_{1,40} = 0.110$, $p = 0.372$).

III. Discussion

The main conclusion from these experiments is that learning to understand noise-vocoded speech is not specific to the trained regional accent, much less to the voice used during training. The experiments demonstrate that participants who are trained to comprehend degraded speech learn information about the degradation itself (i.e., how phonemes are transformed when they are noise-vocoded) rather than learning information that is specific to a given talker or accent. All trained groups had better performance than naïve groups tested with the same materials during the first ten test trials. In Experiments 1 and 2, training with non-native-accented NV speech and training with native-accented NV speech resulted in similar levels of post-training comprehension.

It is particularly striking that the groups who heard different talkers with different accents during training and test (Switch Condition) performed as well during the test as the groups who heard a single talker throughout training and test (Constant condition). The transfer of learning between accents might not have been as complete if the non-native accent were more pronounced, or less familiar. Nevertheless, the evidence that learning generalizes completely from one talker, to a talker who is not only a different person but has a different accent, is relatively strong: performance in the Switch conditions did not differ from that in the Constant conditions at any time point during testing. The Constant conditions defined the greatest possible transfer of learning from training to test (since only the sentences changed).

Note that it is possible that the nature of the training materials might matter to absolute performance levels during training and testing. Some materials (i.e., Voice and Sentence combinations) might be easier to comprehend than others, as seen in Experiment 1, where there was better training performance for native vs. non-native accented speech. There may also be an interaction between training and test materials — such that some training materials 'set listeners up better' for some test materials than for others. However, in Experiment 1, since the training materials were constant across conditions, as were the test materials, this absolute difference canceled out, and the relative differences (how different types of training lead to different levels of test performance) are interpretable.

Hervais-Adelman et al. (2011) previously demonstrated complete generalization to untrained frequency regions: Listeners who heard degraded sentences that had been filtered into one frequency range (50-1406 Hz or 1593-5000 Hz) during training and the other frequency range during the test phase had similar word report scores at test compared to listeners who heard degraded sentences in the same frequency range throughout training and test. In another

experiment in that paper, there was *incomplete* generalization among different carriers used to generate vocoded speech (i.e., sine waves, pulse trains, and noise bands). Listeners who were trained and tested with vocoded speech generated with the same carrier exhibited better performance than those who were trained and tested with speech generated with two different carriers. Taken together, these results suggest that learning to comprehend vocoded speech occurs at a stage of processing at which the stimulus representation has been somewhat abstracted from the acoustic signal but still includes certain acoustic features such as periodicity and noise (which are carrier-specific).

Dahan and Mead (2010), in their study of perceptual learning of single noise-vocoded words, demonstrated only partial generalization between NV consonants in different acoustic contexts (i.e., initial vs. final position within a word), consistent with the idea that, at the level of processing at which learning occurs, context-specific acoustic attributes of the stimulus are still relevant.

When they further examined the learning of context-specific attributes, by examining whether learning generalized to a different talker, Dahan and Mead's (2010) results were inconsistent. In one experiment they observed complete generalization between voices of different genders (that were easy to distinguish after vocoding) but in another experiment they observed incomplete generalization between voices of the same gender (that were difficult to distinguish after vocoding). The authors interpreted this outcome to mean that learning to perceive noise-vocoded speech may involve changes in representations that include information about the particular acoustic properties of the voice but that this voice-specific information is ignored when the listener can easily tell that the trained and untrained voices are different. This could have been a factor in the present study, especially because the listeners heard clear

versions of all stimuli and thus were likely aware that they were listening to a novel talker. Further, while Dahan and Mead (2010) used words as their stimuli, the use of sentences in the present study probably provided additional clues that enabled the listeners to recognize that the test voice was different than the trained voice, despite similarities in the talker characteristics.

In the present experiment it appears that listeners are learning something about the relationship between clear and degraded stimuli—that is, about the stimulus transformation—and that they are able to apply that learning to novel stimuli. When listeners are tested on comprehension of speech in noise (Nygaard et al., 1994; Nygaard & Pisoni, 1998; Johnsrude et al., 2013) or identification of word boundaries (Smith & Hawkins, 2012), performance is almost always better for novel stimuli from the same talker than for stimuli from a new talker. This talker-specific learning contrasts dramatically with what has been observed here. However, the exposure to the training talker was brief in our study: listeners were trained on only 20 sentences, whereas Smith & Hawkins' (2012) listeners heard 288 sentences, and the familiar voice for each listener in Johnsrude et al. (2013) was that of their spouse. Moreover, our training task placed no focus on voice learning, unlike Nygaard et al. (1994) or Nygaard & Pisoni (1998), where listeners' task in their training period was to identify ten voices. These aspects of our design may explain why we did not observe an advantage for the trained talker.

More broadly, learning to understand a systematic distortion like noise-vocoding may differ from perceptual learning of voices in terms of what, exactly, is learned. Here, listeners appear to be learning the rules that govern the transformation from clear to noise-vocoded speech - they appear to be learning the lawful regularities in the transform, which are by definition constant for all types of speech, and not the specific acoustics of phoneme realizations in NV form (which differ substantially between talkers and accents).

The information transmitted in vocoded speech pertains mainly to rhythmic properties: timing and relative loudness. Information about sibilant fricatives, certain other obstruents, and some spectral information about vowel quality is also retained. These rhythmic and vestigial spectral properties can be fitted directly into the language's expected phonotactic, word, and grammatical patterns, and this may be what participants are listening for, regardless of whether the signal is distorted or not. For vocoded speech, an unfamiliar (or less familiar) accent of the listener's native language violates native expectations largely to the extent that the rhythmic properties of the two accents differ. SSBE and Canadian English are rhythmically rather similar, and this may be why learning transferred so well from one to the other. Nevertheless, as is often the case with SSBE, especially that spoken by young people, there were more devoiced syllables in the sentences from the SSBE talker than the CE talker, and devoicing affected a much wider range of words in SSBE than CE (i.e., many function and some content words in SSBE, as opposed to just the function words 'into' and 'to' in CE). These devoiced syllables somewhat disrupt the canonical syllable pattern shared by the two accents. In consequence, the canonical forms may have been less obvious to Canadian listeners unfamiliar with the SSBE forms, and/or there may have been disruption to their interpretation of the rhythm of the surrounding material as well. While devoicing was clearly not disruptive enough to prevent generalization from SSBE to CE or vice versa, it may partially explain why the Canadian listeners performed more poorly with the non-native-accented NV speech than they did with native-accented NV speech.

While our training paradigm produced both learning and generalization, it may not have been optimal. Previous research suggests that presenting multiple talkers during training could be beneficial to learning (Bradlow and Bent, 2008; Baese-Berk, Bradlow, and Wright, 2013). Learning (and generalization) in the present study also may have been limited by ceiling effects:

Listeners in the Naïve group performed equally as well as the trained (Switch and Constant) listeners by second set of test sentences. Thus, there may have been a limit in how much listeners could improve on the training and testing materials that obscured possible differences in generalization between the Switch and Constant groups. If so, this limit could have been due to acoustic or cognitive factors.

IV. Concluding Remarks

The main results are clear: listeners generalized completely from one talker (native or nonnative accent) during training, to a different talker (native or nonnative accent) during testing. Thus, learning of noise-vocoded speech does not appear to be specific to a particular talker or accent. Because this study involved learning to understand spectrally degraded (noise-vocoded) speech, the results suggest that people with cochlear implants, who experience spectrally degraded speech through their prostheses, may not be disadvantaged if they learn to understand speech through their implant by listening much of the time to just one other talker (e.g., their spouse). The evidence suggests that, at least insofar as the accents are rhythmically similar, such training will allow the listener to generalize completely to a wider range of voices and accents both in the laboratory and in the real world.

Acknowledgments

A. Chau, V. Cheung, A. Krishna, J. Riley and T. Viaznikova helped with data collection and scoring. L. Bailey assisted with stimulus generation and programming. This work was supported by the Canadian Natural Sciences and Engineering Research Council (NSERC) through a Discovery Grant and EWR Steacie supplement to ISJ. JJH received support from

NSERC and from the F.V. Hunt Post-Doctoral Fellowship from the Acoustical Society of America.

References

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. http://doi.org/10.1016/j.cognition.2007.04.005.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America, 133*(3), EL174-EL180. http://doi.org/10.1121/1.4789864.

Dahan, D., & Mead, R. L. (2010). Context-Conditioned Generalization in Adaption to Distorted Speech. *Journal of Experimental Psychology. Human Perception and Performance, 36*(3): 704-728. http://doi.org/10.1037/a0017449.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222–241. http://doi.org/10.1037/0096-3445.134.2.222

Fu, Q.-J., & Galvin, J. J. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *The Journal of the Acoustical Society of America*, *113*(2), 1065–1072. http://doi.org/10.1121/1.1537708

Greenwood, D. D. (1990). A cochlear frequency-position function for several species--29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–605. http://doi.org/10.1121/1.399052

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(2), 460–474. http://doi.org/10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(1), 283–295. http://doi.org/10.1037/a0020772

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: voice familiarity AIDS speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. http://doi.org/10.1177/0956797613482467

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376. http://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42-46. http://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269. http://doi.org/10.1093/cercor/bhi009

Shannon, R. V, Zeng, F.-G. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science (New York, N.Y.)*, *270*(5234), 303–4. http://doi.org/10.1126/science.270.5234.303

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics*, *40*(2), 213–233. http://doi.org/10.1016/j.wocn.2011.11.003

Studebaker, G. a. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research*, *28* (September 1985), 455–462. http://doi.org/10.1044/jshr.2803.455

Wayne, R. V, & Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology. Applied*, *18*(4), 419–35. http://doi.org/10.1037/a0031042

Collected figure captions

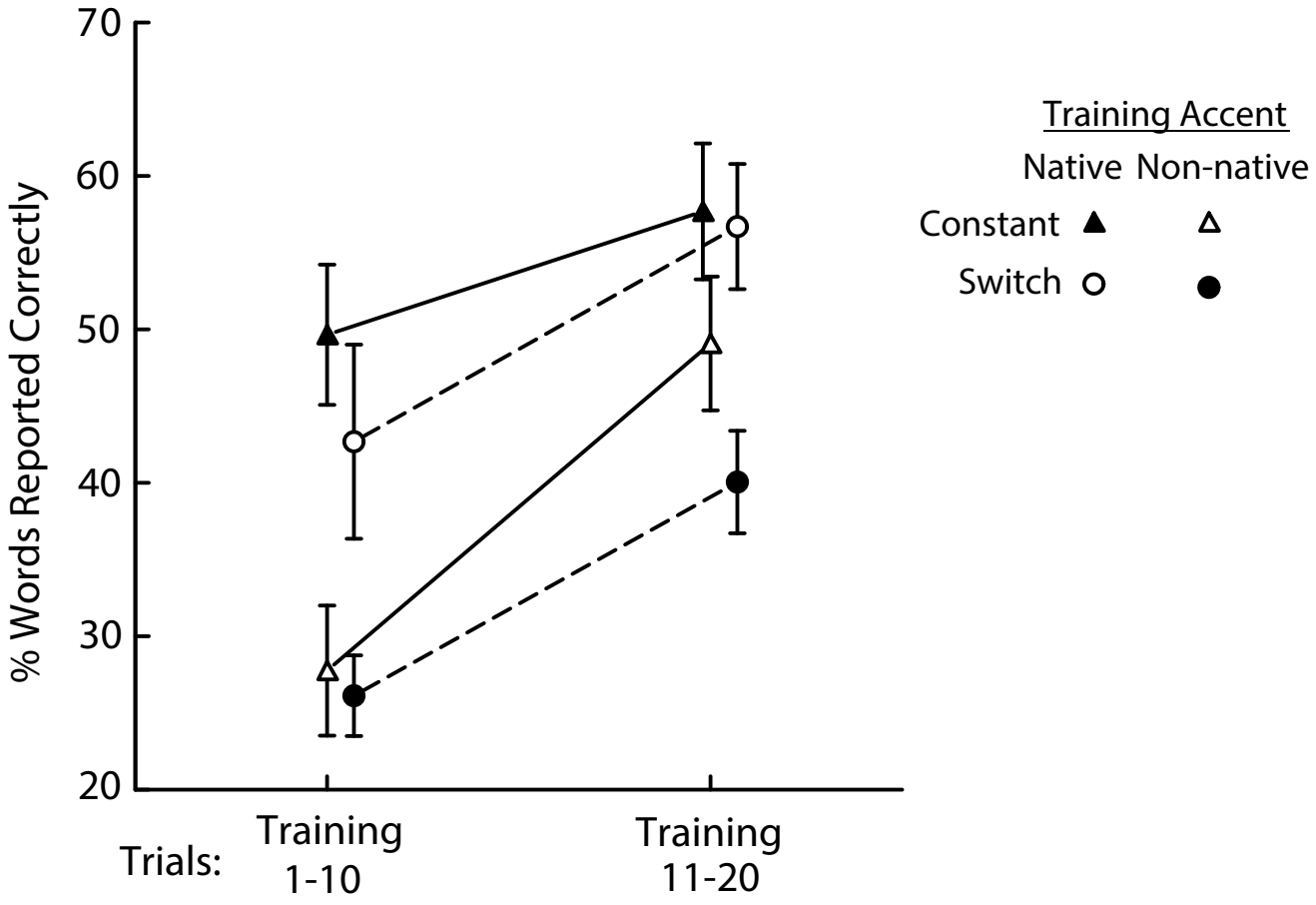
Figure 1. Mean word-report scores for each ten-sentence bin during the training for Experiment 1. Data are shown separately for participants who heard the same accent during training and testing (Constant condition; triangles), and those who heard a different accent and voice during training than during testing (Switch condition; circles). Data are also shown separately for listeners who heard their native (Canadian) accent during the training (left column of key) and those who heard the non-native (SSBE) accent during the training (right column of key). Filled symbols represent listeners who eventually heard their native accent during the post-training test while open symbols represent those who eventually heard the non-native (SSBE) accent during the post-training test (open symbols). Error bars indicate +/- one standard error of the mean.

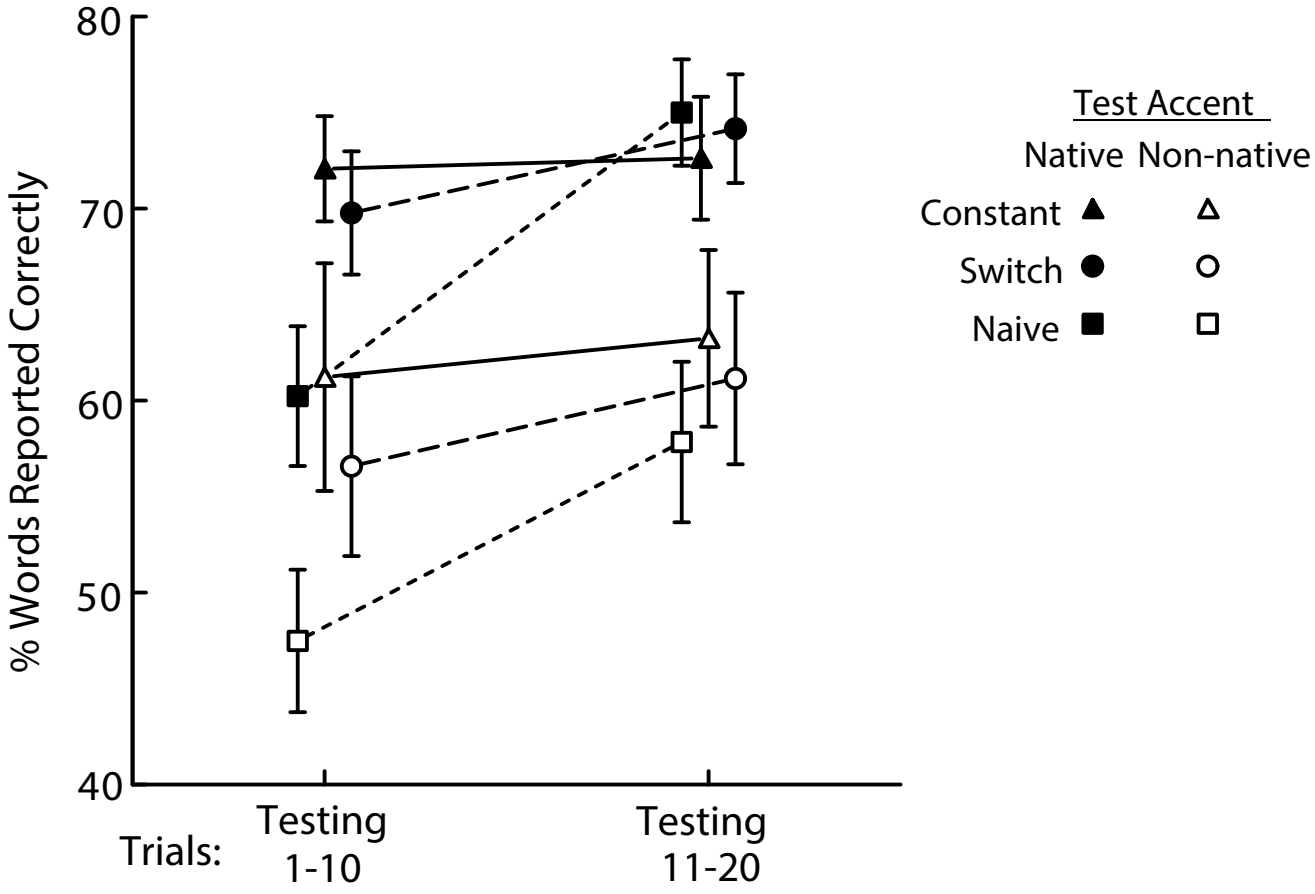
Figure 2. Mean word-report scores for each ten-sentence bin during the post-training test for Experiment 1. Data are shown separately for participants in the Naïve condition (squares), those

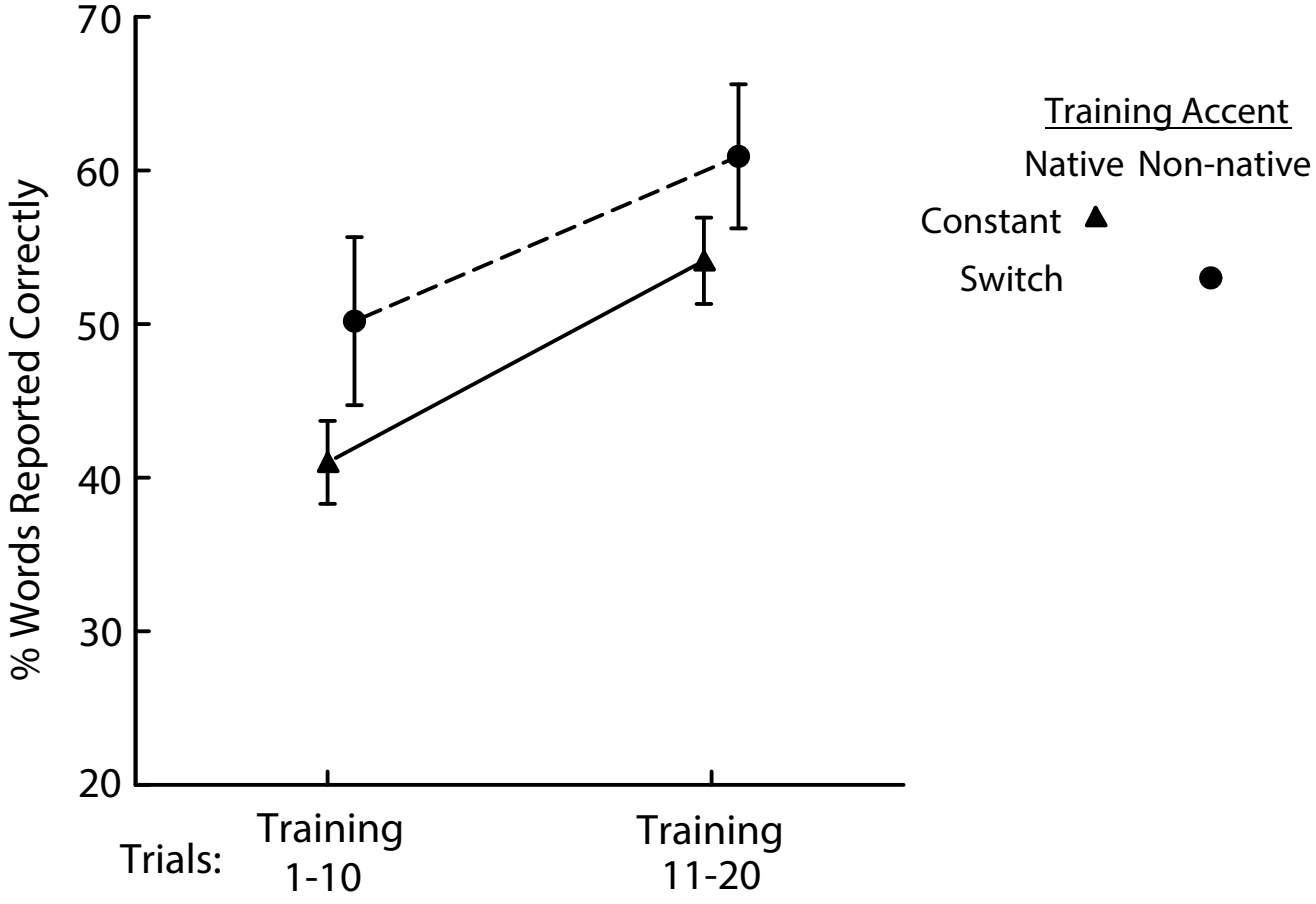
who heard the same accent during training and testing (Constant condition; triangles), and those who heard a different accent and voice during training than during testing (Switch condition; circles). Data are also shown separately for listeners who heard their native (Canadian) accent during the test (filled symbols) and those who heard the non-native (SSBE) accent during the test (open symbols). Error bars indicate +/- one standard error of the mean.

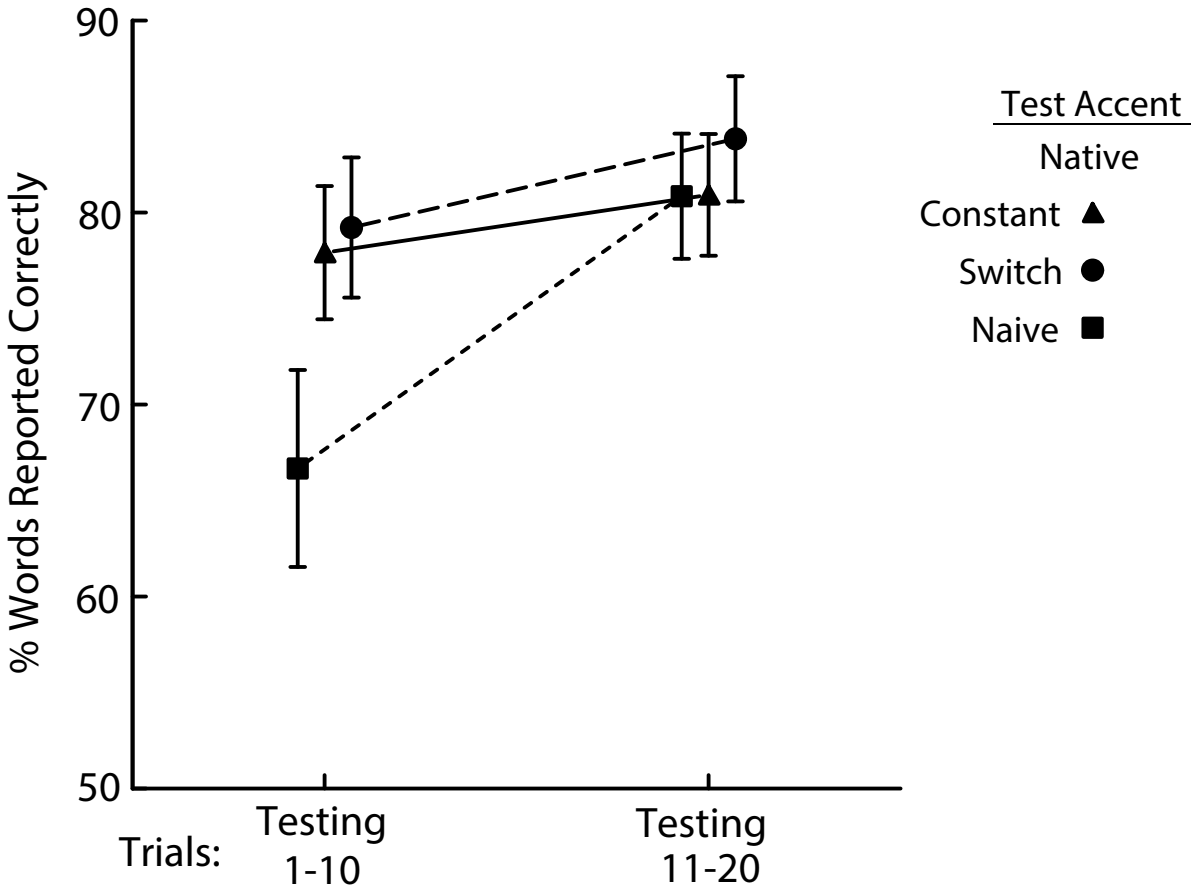
Figure 3. Mean word-report scores for each ten-sentence bin during the training for Experiment 2. Data are shown separately for participants who heard the same (native, SSBE) accent during training and testing (Constant condition; triangles), and those who heard a non-native (Canadian) accent during training and a native (SSBE) voice during testing (Switch condition; circles). Error bars indicate +/- one standard error of the mean.

Figure 4. Mean word-report scores for each ten-sentence bin during the post-training test for Experiment 2. Data are shown separately for participants in the Naïve condition (squares), those who heard the same accent during training and testing (Constant condition; triangles), and those who heard a different accent and voice during training than during testing (Switch condition; circles). All listeners heard their native (SSBE) accent during the test. Error bars indicate +/- one standard error of the mean.


70
60
50
40
30
20
% Words Reported Correctly
Trials:
Training
1-10
Training
11-20
Training Accent
Native
Non-native
Constant
Switch


80
70
60
50
40
% Words Reported Correctly
Trials:
Testing
1-10
Testing
11-20
Test Accent
Native
Non-native
Constant
Switch
Naive


70
60
50
40
30
20
% Words Reported Correctly
Trials:
Training 1-10
Training 11-20
Training Accent
Native
Non-native
Constant
Switch


90
80
70
60
50
% Words Reported Correctly
Trials:
Testing
1-10
Testing
11-20
Test Accent
Native
Constant
Switch
Naive