Article

Generalization of perceptual learning of vocoded speech

HERVAIS-ADELMAN, Alexis, *et al*.

Abstract

Recent work demonstrates that learning to understand noise-vocoded (NV) speech alters sublexical perceptual processes but is enhanced by the simultaneous provision of higher-level, phonological, but not lexical content (Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008), consistent with top-down learning (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman et al., 2008). Here, we investigate whether training listeners with specific types of NV speech improves intelligibility of vocoded speech with different acoustic characteristics. Transfer of perceptual learning would provide evidence for abstraction from variable properties of the speech input. In Experiment 1, we demonstrate that learning of NV speech in one frequency region generalizes to an untrained frequency region. In Experiment 2, we assessed generalization among three carrier signals used to create NV speech: noise bands, pulse trains, and sine waves. Stimuli created using these three carriers possess the same slow, time-varying amplitude information and are equated for naïve intelligibility but differ in their temporal fine [...]

Reference

HERVAIS-ADELMAN, Alexis, *et al*. Generalization of perceptual learning of vocoded speech. *Journal of experimental psychology. Human perception and performance*, 2011, vol. 37, no. 1, p. 283-295

DOI : 10.1037/a0020772
PMID : 21077718

Journal of Experimental Psychology:
Human Perception and Performance
2011, Vol. 37, No. 1, 283–295

0096-1523/10/$12.00 DOI: 10.1037/a0020772

Generalization of Perceptual Learning of Vocoded Speech

Alexis G. Hervais-Adelman
Medical Research Council, Cognition and Brain Sciences Unit, United Kingdom and University of Geneva

Matthew H. Davis
Medical Research Council, Cognition and Brain Sciences Unit, United Kingdom

Ingrid S. Johnsrude
Queen's University

Karen J. Taylor
Medical Research Council, Cognition and Brain Sciences Unit, United Kingdom and University of California–Davis

Robert P. Carlyon
Medical Research Council, Cognition and Brain Sciences Unit, United Kingdom

Recent work demonstrates that learning to understand noise-vocoded (NV) speech alters sublexical perceptual processes but is enhanced by the simultaneous provision of higher-level, phonological, but not lexical content (Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008), consistent with top-down learning (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman et al., 2008). Here, we investigate whether training listeners with specific types of NV speech improves intelligibility of vocoded speech with different acoustic characteristics. Transfer of perceptual learning would provide evidence for abstraction from variable properties of the speech input. In Experiment 1, we demonstrate that learning of NV speech in one frequency region generalizes to an untrained frequency region. In Experiment 2, we assessed generalization among three carrier signals used to create NV speech: noise bands, pulse trains, and sine waves. Stimuli created using these three carriers possess the same slow, time-varying amplitude information and are equated for naïve intelligibility but differ in their temporal fine structure. Perceptual learning generalized partially, but not completely, among different carrier signals. These results delimit the functional and neural locus of perceptual learning of vocoded speech. Generalization across frequency regions suggests that learning occurs at a stage of processing at which some abstraction from the physical signal has occurred, while incomplete transfer across carriers indicates that learning occurs at a stage of processing that is sensitive to acoustic features critical for speech perception (e.g., noise, periodicity).

Keywords: speech perception, learning, adaptation, language, generalization

The human speech perception system is able to cope with wide variation in the sounds of speech. On a day-to-day basis we understand speech despite variability attributable to the size and gender of the talker (e.g., Peterson & Barney, 1952; or more recently Smith & Patterson, 2005), different accents (e.g. Clarke & Garrett, 2004), or changes to speech rate (e.g., Altmann & Young, 1993). In laboratory situations, speech perception is remarkably robust even to artificial and more severe distortions and degradations of the signal (e.g., Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995).

In many situations, perceptual learning is crucial to successful perception of degraded speech. Goldstone (1998) defines perceptual learning as "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment." Exposure to distorted or degraded speech in an appropriate learning situation allows the perceptual system to adapt so as to process similarly distorted input more effectively in future. For instance, Norris, McQueen, and Cutler (2003) showed that exposure to speech containing ambiguous fricative phonemes that are disambiguated by lexical context alters the categorization of similar sounds in a subsequent test session.

Here, we explore perceptual learning of vocoded speech, an artificial manipulation that removes much of the fine spectral detail from speech while leaving slow amplitude fluctuations intact (Shannon et al., 1995). Vocoding involves separating a signal into nonoverlapping frequency bands and extracting the time-varying amplitude envelope from each of these bands. The extracted amplitude envelopes are then used to modulate a carrier signal which

This article was published Online First November 15, 2010.

Alexis G. Hervais-Adelman, Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, United Kingdom and Functional Brain Mapping Lab, University of Geneva; Matthew H. Davis and Robert P. Carlyon, Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, United Kingdom; Ingrid S. Johnsrude, Department of Psychology, Queen's University; Karen J. Taylor, Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, United Kingdom and Department of Psychology, University of California–Davis.

This work was supported by a United Kingdom Medical Research Council PhD studentship to Alexis Hervais-Adelman, and by Medical Research Council funding of the Cognition and Brain Sciences Unit (U.1055.04.013.01.01, Robert P. Carlyon and Matthew H. Davis), by a grant from the Leverhulme Trust (Karen J. Taylor). and by the Canada Research Chairs Programme (Ingrid S. Johnsrude). We are grateful to Chris Darwin and two anonymous reviewers for their very helpful comments and suggestions.

Correspondence concerning this article should be addressed to Matthew H. Davis, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom. E-mail: matt.davis@mrc-cbu.cam.ac.uk

replaces the fine structure of the original signal. Vocoded speech is frequently used as a simulation of sound transduced by a cochlear implant, and investigators have explored how different parameters of this manipulation affect speech intelligibility (Faulkner, Rosen, & Smith, 2000; Loizou, Dorman, & Tu, 1999; Shannon et al., 1995). For instance, variations in the number and spacing of the frequency bands used in creating vocoded speech affect intelligibility in normally hearing listeners in a manner that resembles the effect of changing the number and placement of electrodes in cochlear implant users (Fu & Galvin, 2003; Rosen, Faulkner, & Wilkinson, 1999; Shannon et al., 1995).

Vocoded speech has been shown to be subject to perceptual learning; it becomes more comprehensible if listeners are trained to understand it (e.g., Davis et al., 2005; Shannon et al., 1995). We have previously demonstrated that word report scores for vocoded sentences improves from around 0% to 70% correct following exposure to just 30 distorted sentences (Davis et al., 2005). We also showed that foreknowledge of the identity of noise-vocoded (NV) sentences (the clear spoken or written form of the sentence) enhances perceptual learning.

One key finding that has emerged from our previous work on sentences (Davis et al., 2005) and single vocoded words (Hervais-Adelman, Davis, Johnsrude, Carlyon, 2008) is that perceptual learning of vocoded speech generalizes to untrained words. This finding demonstrates that adaptation does not merely involve rote learning of distorted words, or enhanced guessing of likely sentence content, but rather that perceptual adaptation is general to multiple lexical items. This constrains the choice of functional levels at which changes underlying perceptual learning must occur, suggesting that prelexical processing of speech has been modified so as to more effectively represent the critical information that is found in distorted or degraded speech input. Generalization between lexical items has similarly been observed for accented speech (Clarke & Garrett, 2004), synthetic speech (Fenn, Nusbaum, & Margoliash, 2003), and speech containing ambiguous phonemes (McQueen, Norris, & Cutler, 2006).

Here, we assess the degree to which perceptual learning of vocoded speech generalizes from one set of acoustic characteristics to another, to further constrain the locus of learning. This test of acoustic generalization is an application of the "psychoanatomical" method, first described by Julesz (1971) in which knowledge of the underlying anatomy of a cognitive process is used to infer the neural locus of a particular behavior (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002). For instance, in Julesz's work using random-dot stereograms, it was shown that certain visual illusions (such as fluctuations in perception of a Necker cube) remain effective when form information is conveyed only by binocular disparity. Hence, neural mechanisms for generating this illusion must rely on a stage of visual processing after input from the two eyes is combined in the lateral geniculate body of the thalamus. A similar method has been applied in studies of visual perceptual learning (Ahissar & Hochstein, 2004): perceptual learning of orientation discrimination is specific to the region of the retina in which stimuli were presented during training, indicating that perceptual learning involves changes to retinotopically organized processing stages early on in the cortical visual pathway.

In investigations of auditory perceptual learning, a key experimental question concerns the frequency specificity of perceptual learning. The auditory system is organized tonotopically at least up to belt auditory cortex (Kaas & Hackett, 2000; Rauschecker & Tian, 2000). Hence, learning effects that are specific to a trained frequency (i.e., a particular region of the cochlea) would result from processes relying on tonotopically organized levels of the pathway; i.e., belt auditory cortex, or below. Frequency-discrimination training has been shown to generalize to the untrained ear and across frequency (Roth, Amir, Alaluf, Buchsenspanner, & Kishon-Rabin, 2003), suggesting that improvements in frequency discrimination might result from modifications to a part of the auditory system that is not primarily organized according to frequency (see also Demany & Semal, 2002; Grimault, Micheyl, Carlyon, Bacon, & Collet, 2003). In contrast, other forms of auditory perceptual learning have been found to be specific to trained frequency regions. For example, Fitzgerald and Wright (2000) found that improvements in amplitude-modulation (AM) rate discrimination are specific to the frequency of the carrier signal used during training. Such a result suggests that perceptual learning of AM rate description modifies auditory processes that are organized according to frequency. Given the importance of AM information in the perception of vocoded speech, we might expect that perceptual learning of vocoded speech would be similarly specific to the frequency of the carrier signal.

A previous study (Fu & Galvin, 2003) has examined whether perceptual learning generalizes across different forms of frequency-shifted vocoded speech. In frequency shifted vocoded speech, the amplitude envelope of speech originating in a certain frequency region is used to modulate a carrier signal in a different frequency region. Such a manipulation is intended to simulate cochlear implants which present low-frequency speech information to regions of the cochlea that would normally respond to higher-frequency energy, attributable to necessarily incomplete insertion of the electrode array into the cochlea. This study showed that, compared with unshifted vocoded speech, frequency-shifted vocoded speech is both more difficult to comprehend and is learned more slowly (see Rosen et al., 1999, for more information on the impact of these pitch-shift simulations). In the studies by Fu and Galvin (2003), listeners were trained to understand one form of frequency shifted vocoded speech and then tested with different frequency shifts. Although listeners' performance improved for trained stimuli, the improvement did not generalize to other, untrained frequency shifts, perhaps implicating local adaptation, specific to the trained frequency. However, this could reflect either perceptual learning of vocoded speech in specific frequency regions, or perceptual learning of a particular envelope-carrier transformation. Evidence from the perception of speech produced in a helium-rich environment (Belcher & Hatlestad, 1983; Morrow, 1971) supports the proposal that transposing speech formants creates a substantial additional obstacle to comprehension even without the loss of spectral detail produced by vocoding. Therefore, different results may be obtained when using an unshifted vocoding manipulation.

We present two experiments that assess the degree of acoustic generalization of perceptual learning of vocoded speech. The first study tests for generalization in the frequency domain by assessing whether listeners trained on unshifted vocoded speech presented in one frequency range show better-than-naïve performance when tested on an otherwise identical manipulation presented in a nonoverlapping frequency range. If perceptual learning depends on enhanced sensitivity to the trained frequency region, we might

expect perceptual learning to be frequency specific (cf. Fitzgerald & Wright, 2000). A second study used identical frequency ranges during training and testing, but tested for generalization of learning among vocoders with three different carrier signals (noise, sine waves, and pulse trains). If perception of vocoded speech depends on enhanced sensitivity to amplitude modulations, learning should generalize among the three carrier signals, which sound different but which all contain identical AM modulation in each frequency range.

Experiment 1: Generalization Between Frequency Ranges

In this experiment, we investigate whether perceptual learning of vocoded speech generalizes over different frequency regions. That is, are listeners who are trained to perceive vocoded speech that has been high-pass filtered also able to perceive low-pass filtered vocoded speech (i.e., in a nonoverlapping frequency range) at better than naïve levels of performance (and vice-versa for listeners trained with low-pass vocoded speech)? As in other explorations of perceptual learning, a failure of generalization would suggest that tonotopically organized stages of processing are modified by perceptual learning. For consistency with the previous study using frequency-shifted vocoded speech (Fu & Galvin, 2003), all stimuli were created by modulating a noise carrier with speech amplitude envelopes (NV speech).

Method

Participants. Forty-eight listeners from the Medical Research Council Cognition and Brain Sciences volunteer panel took part in this experiment (16 men, 44 right-handed, average age = 21 years and 7 months). Participants self-reported as having no history of hearing impairment or dyslexia.

Materials. Forty declarative English sentences were assigned to one of four sets of 10 (A, B, C, D) matched for number of words per sentence (range = 6 to 13 words, $M = 8.7$, $SD = 1.93$), duration ($M = 2.03$ s, $SD = 0.441$) naturalness ($M = 6.96$, $SD = 0.557$), and imageability ($M = 6.6$, $SD = 1.05$). The sentences were originally recorded for use as unambiguous controls by Rodd, Davis, and Johnsrude (2005) and naturalness (how likely each sentence is to be used in natural language) and imageability ratings (the ease with which sentences arouse mental images) were obtained in that study. Sentences are listed in the Appendix. The sentences were spoken by a female speaker of Southern British English, and recorded onto digital audio tape at a sampling rate of 48 kHz, digitized, and downsampled to 22.1 kHz using a desktop computer.

Each sentence was filtered into two frequency bands: 50 Hz–1,406 Hz (low) and 1,593 Hz–5,000 Hz (high), using low-pass and high-pass brickwall filters, respectively; these two regions were separated by one equivalent rectangular bandwidth (ERB; calculated on the basis of Glasberg & Moore, 1990), to minimize any possibility of low-pass and high-pass stimuli activating overlapping regions of the cochlea. The two frequency ranges were selected because they yielded equally intelligible results when a separate group of pilot subjects were asked to rate the intelligibility of speech filtered into various passbands. The approximately equal intelligibility of the chosen bands is consistent with the predictions of the ANSI method for calculation of the speech intelligibility index (ANSI, 1997).

The sentences were vocoded using the procedure described by Shannon and colleagues (Shannon et al., 1995) as implemented by Deeks and Carlyon (2004) in Matlab (The Mathworks, Inc., Natick, MA). The low-pass and high-pass sentences were first filtered into quasi-logarithmically spaced frequency bands. Contiguous bandpass filters were constructed in the frequency domain. For the low-pass sentences, six frequency bands were used: corner frequencies of the filters (3dB down from the peak of the passband) were at 50 Hz, 135 Hz, 252 Hz, 416 Hz, 644 Hz, 963 Hz, and 1,406 Hz. For the high-pass sentences, six frequency bands were initially used, but because of experimenter error a final stage of low-pass filtering was implemented with a cutoff at 5 kHz. This left five frequency bands for the vocoding, whose corner frequencies (3dB down) were at 1,939 Hz, 2,353 Hz, 2,848 Hz, 3,441 Hz, 4,151 Hz, and 5,000 Hz. In all cases, filters had a roll-off of 48dB/octave. These cutoff frequencies were chosen to simulate equal distances along the basilar membrane (based on Greenwood, 1990). The amplitude envelope from each band was extracted by half-wave rectifying the output of the corresponding bandpass filter, and passing it through a second-order low-pass Butterworth filter with a cutoff of 30 Hz. The resulting set of amplitude envelopes was then applied to bandpass filtered noise in the same frequency ranges as the source. The modulated noise carriers were finally recombined to produce the distorted sentences. Figure 1 shows the waveform and spectrogram of an unprocessed sentence, and the low- and high-pass vocoded versions of this sentence.

Design and procedure. Participants were divided into four groups of 12 listeners: two "Transfer" groups that received 20 sentences filtered into the low frequency range, followed by 20 sentences filtered into the high frequency range (Low-High group) or vice versa (High-Low group), and two "No Transfer" groups that received the same form of vocoded speech for all 40 sentences (either low- or high-frequency vocoded speech throughout). The first 20 sentences will be referred to as "naïve," and the next 20 as "switch" (in the conditions testing generalization of learning, when the manipulation type changed after 20 trials) or "no-switch" (in the conditions in which the form of speech manipulation did not change across the 40 items). The four matched groups of 10 sentences were counterbalanced across conditions (equal numbers of participants in each condition being presented with sentence groups in the order ABCD, BADC, CDAB, DCBA). Sentences within a group were always presented in the same order. Using this counterbalancing ensured that each sentence group appeared equally frequently as the first or second stimulus block in all conditions, naïve, switch, or no-switch. Thus, all comparisons between conditions include sentence groups A through D paired in all possible orders. Stimuli were presented over Sennheiser HD250 headphones through a QED headphone amplifier, from a desktop PC fitted with a Soundblaster Live sound card. Participants were asked to listen carefully to each stimulus and to write down whatever they could understand. Each sentence was preceded by a warning tone and followed by a 25-s silent period during which listeners were able to write down the words they understood. After they had finished writing, participants received "feedback," which consisted of the same sentence presented as clear speech (C), and then in its vocoded (degraded) form again (D). This "DCD" presentation order has been shown to be an effective form of

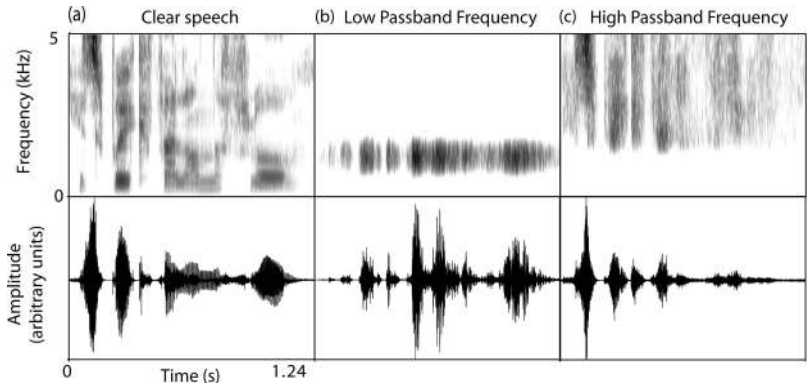

Figure 1. Waveform and spectrograms of the sentence "His wig fell on the floor"; (a) The original sentence, (b) low-passband vocoded, (c) high-passband vocoded sentence.

training, in that it leads to optimal perceptual learning of vocoded speech (Davis et al., 2005). Davis et al. (2005) have shown that after 20 sentences of exposure to unfiltered (i.e., spanning the range 50–5,000 Hz) NV speech, with this feedback, listeners' performance improves significantly, by an average of almost 20% of words reported correctly. Listeners' performance was scored as percentage of words reported correctly for each sentence. Homophones of target words were scored as correct, but incorrectly reported inflected forms (for example "jump" instead of "jumped") were not.

Results

The results are shown in Figure 2. We analyzed high- and low-passband speech separately. Data were averaged across each block of 10 sentences over all participants. To establish whether subjects learned, we compared naive performance with performance on the second two blocks of sentences in the "no-switch" groups. We conducted a repeated-measures analysis of variance (ANOVA) with block (two levels: first or second set of 10 sentences) and condition (naïve vs. no-switch) as within-participants factors. Only the 12 participants in the no-switch conditions were included in these analyses. An additional between-participants dummy variable was included, coding for the order in which the four sentence groups (ABCD) were presented to each participant. Effects of this variable will not be reported (as suggested by Pollatsek and Well, 1995).

To determine whether there was generalization of learning from one frequency region to another, we compared performance on the switch sentences with naïve and no-switch sentences in turn. Data were entered into two repeated-measures ANOVAs, with condition as a 2-level between-subjects factor and block as a 2-level within-subjects factor (first or second group of 10 sentences). In the comparisons with naïve performance, the 24 data-sets from the naïve conditions were compared with the 12 datasets from the switch conditions. The results of the analyses are presented below. When performing analyses, data were averaged over participants but not over items (cf. Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999). We present one-tailed p values, because we have directional hypotheses: we predict that performance in the no-switch groups will be equivalent to or better than naïve, that performance in the switch conditions will be equivalent to or better than naïve and equivalent to or worse than the no-switch conditions. In additional analyses not reported here, we have shown that there are no conditions in which performance in a switch or no-switch condition is significantly worse than naïve. We elected not to report tests of within condition differences (i.e., between sentences 1–10 and 11–20 or between sentences 21–30

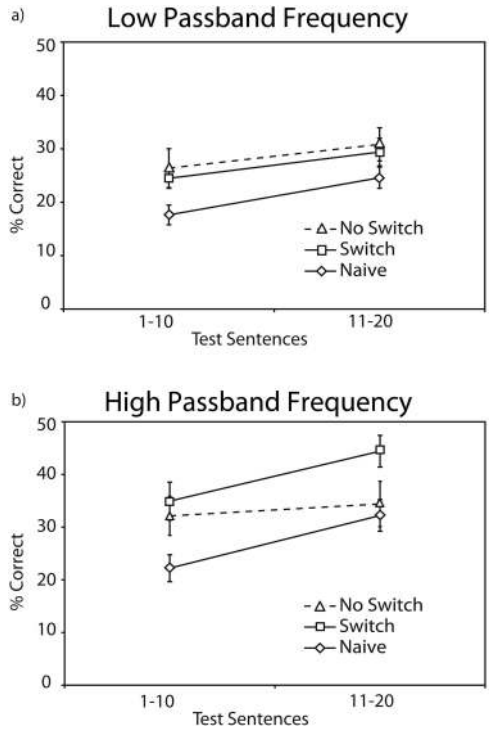

Figure 2. Results of Experiment 1, showing performance averaged over participants over 10-sentence blocks. Test sentences 1–10 and 11–20 indicate the first and second 10-sentence blocks in each condition Error bars represent ± 1 *SEM*. (a) Performance on low-passband vocoded speech. (b) Performance on high-passband vocoded speech. Naïve, $N = 24$; no-switch and switch, $N = 12$.

and 31–40) because comparisons between adjacent blocks of ten sentences have less statistical power to detect the expected effects of perceptual learning. Because Davis et al. (2005) have already demonstrated significant improvements in NV speech comprehension over 20 sentences with CD feedback, we focus instead on comparisons between naïve and no-switch conditions where perceptual learning can be assessed over two blocks of twenty sentences (1–20, 31–40).

Low-passband vocoded speech. Performance in the no-switch condition was significantly better than naïve performance, $F(1,8) = 22.988$, $p = .001$, partial $\eta^2 = 0.742$, indicating that learning has taken place. Performance in the switch condition was significantly better than naïve, $F(1,28) = 5.210$, $p = .015$, partial $\eta^2 = 0.157$. There was no significant difference in performance between the no-switch and switch conditions, $F(1,16) = 0.698$, $p = .208$, partial $\eta^2 = 0.042$). Training with high-passband vocoded speech improves performance on low-passband vocoded speech.

High-passband vocoded speech. Performance in the no-switch condition was significantly better than naïve performance, $F(1,8) = 12.974$, $p = .003$, partial $\eta^2 = 0.619$, showing that learning occurred. Performance in the switch condition was significantly better than naïve, $F(1,28) = 10.606$, $p = .001$, partial $\eta^2 = 0.275$. There was no difference between switched and no-switch trials, $F(1,16) = 2.382$, $p = .071$, partial $\eta^2 = 0.130$—indeed, even this marginally significant difference was in the opposite direction to that predicted (i.e., participants in the switch condition perform numerically better than those in the non-switch condition, mean(switch) = 40.1%, mean(non-switch) = 34.2%). We had no specific prediction concerning interactions between sentence block and condition. The interaction apparent in Figure 2b might suggest that participants in the switch group showed a greater improvement in performance after changing to high-passband vocoded speech. However, this interaction was also only marginally significant, $F(1,16) = 3.477$, two-tailed $p = .081$, partial $\eta^2 = 0.179$. Despite these paradoxical suggestions of improved performance in the switch condition, it should be clear that training with low-passband vocoded speech improved performance on high-passband vocoded speech at least as much as training on high-passband vocoded speech itself.

Discussion

These results indicate that learning bandpass filtered NV speech generalizes completely between different frequency regions. The performance in both the switch groups was statistically indistinguishable from the performance of listeners who had previously received specific training on vocoded speech in the same frequency region (low-passband or high-passband). Hence, perceptual learning of vocoded speech applies equivalently to stimuli that are low- and high-passband filtered. The most plausible means by which this result could arise is if learning leads to modifications of perceptual representations that are not frequency specific. However, it is also possible that activity at a level of the system that is not primarily organized by frequency could cause changes to lower-level perceptual processing that is organized by frequency. For example, a frequency-general mechanism may be in operation which could exercise top-down feedback on multiple, frequency-specific processes. Nonetheless, we would suggest that our results implicate a perceptual learning process in which nonfrequency-specific perceptual representations play a critical role. Given that the primary organizing principle of the ascending auditory system is frequency, our results implicate regions beyond primary auditory cortex as the most likely locus of perceptual learning.

To further explore the exact nature of the representations involved in perceptual learning, we now turn to an investigation of generalization between vocoded speech created using different carrier signals. Although the three carrier signals that we use are acoustically very different, differing in terms of their temporal fine structure, they maintain the slowly changing amplitude envelope of speech. These experiments will determine whether learning is occurring at a level of processing that encodes the envelope or the acoustic fine structure.

Experiment 2: Generalization Between Different Carrier Signals

Smith, Delgutte, and Oxenham (2002) dichotomize speech as consisting of two separate streams of information: the slowly varying amplitude envelope (preserved in a vocoder) and the rapidly varying fine structure. A decomposition based on the Hilbert transform allows creation of "chimeric" stimuli, produced using the fine-structure of one stimulus and the envelope of another. For these chimeras, the acoustic domain in which speech content is conveyed depends upon the number of frequency bands into which the signal is divided. Consistent with the results obtained for similar, NV stimuli (Shannon et al., 1995), if a large number of frequency bands (four or more) are used, then the amplitude envelope of a speech-noise chimera conveys most speech information. In contrast, for chimeras created with very few frequency bands (two or fewer), it is the fine structure that conveys most speech content. Although these results demonstrate the redundant nature of the information contained in natural speech, the two kinds of information are not interchangeable: Lorenzi and colleagues (Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006) have recently demonstrated that whereas normally hearing listeners can understand speech composed either of fine-structure information or envelope information, hearing impairment produces a substantial decline in the ability to perceive speech from fine-structure alone. This result suggests that amplitude envelope cues provide more robust information for speech perception. Smith and colleagues (2002) also observed that fine-structure information was more important than envelope information for sound localization.

In the present study, we explored whether perceptual learning of vocoded speech generalizes among three different carrier signals—noise, pulse trains, and sine waves. We can characterize these three carrier signals as providing either a broadband (noise, pulse trains), or a narrowband (sine waves) structure. As an alternative, we can contrast carrier signals that include periodicity (pulse trains, sine waves) with one that does not (noise). Using different carrier signals in the vocoding process produces a dramatic change in the temporal fine structure of vocoded speech, but essentially preserves the amplitude envelope from the original signal, setting up a dichotomy between envelope and fine-structure in much the same way as Smith et al. (2002). Such a manipulation provides a natural contrast with the high- and low-passband vocoded speech in Experiment 1 which used the same type of carrier, applied to non-overlapping frequency bands. If the generalization

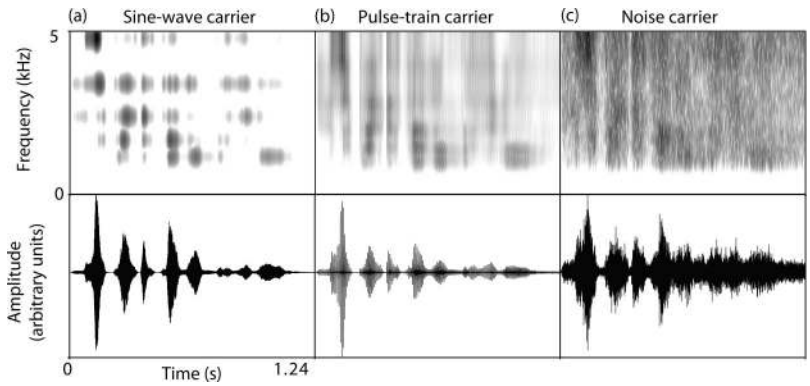

Figure 3. Waveform and spectrograms of the sentence "His wig fell on the floor." (a) sine-wave vocoded, (b) pulse-train vocoded, (c) noise-vocoded vocoded. The original waveform and spectrogram are shown in Figure 1a.

that we observed in Experiment 1 is because of the common carrier signal used in low- and high-passband NV speech, we should observe a lack of generalization among our three carrier signals. On the other hand, if generalization can occur based on utilization of amplitude envelopes, we should observe generalization among differing carrier signals.

Method

Participants. A total of 108 listeners from the Cognition and Brain Sciences Unit volunteer panel took part in the experiment (mean age = 21 years and 5 months, 49 men, 94 right-handed). Participants reported having no history of hearing impairment or dyslexia.

Materials. The same 40 clear sentences were used as in Experiment 1. They were vocoded using a similar procedure to that described in Experiment 1, though three different types of carrier signals were used to vocode each sentence: noise bands (yielding NV speech, NV; as in Experiment 1), sine waves (sine-wave vocoded speech, SW) and pulse trains (pulse-train vocoded speech, PT). The pulse trains were composed of an alternating-phase harmonic complex with a fundamental frequency (F0) of 70 Hz. Summing the components of the harmonic complex in alternating phase produced a pulse repetition rate of 140 pps[1] to remove place-of-excitation cues (which would provide listeners with potentially distracting resolvable pitch information unavailable in the other conditions) when using the pulse-train carriers, the original sentences were high-pass filtered above 937 Hz using an eighth-order Butterworth filter with roll-off of 48 dB/octave (cf. Deeks & Carlyon, 2004) for all conditions.

The high-pass filtered sentences were vocoded using the procedure described by Shannon (Shannon et al., 1995) using a Matlab (The Mathworks, Inc., Natick, MA) algorithm created by Deeks and Carlyon (2004), as in Experiment 1. The sentences were first filtered into 6 frequency bands from 937 Hz–5,000 Hz. Amplitude envelopes were then extracted from 6 passbands, which were 3 dB down at 937 Hz, 1,260 Hz, 1,679 Hz, 2,220 Hz, 2,921 Hz, 3,828 Hz and 5,000 Hz, with a roll-off of 48 dB/octave, simulating equal distances along the basilar membrane (based on Greenwood's equation, 1990). Extracted envelopes were half-wave rectified and low-pass filtered below 30 Hz with an eight-order low-pass Butterworth filter. The resulting sets of six envelopes were applied to each of the three carrier signals: bandpass filtered noise in the same frequency ranges as the source, sine waves with a frequency of the arithmetic midpoint of the source band, and filtered pulse trains. The modulated bands were finally recombined to produce three sets of distorted sentences. It should be noted that the SW condition differs from the sine-wave speech investigated by Remez et al. (e.g., Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez et al., 1981), which is produced using sinusoids that track the formant frequencies and amplitudes of speech to be synthesized.

Informal listening and collection of pilot report score data revealed that the three different types of vocoding were not equally difficult to comprehend, with sine-wave vocoded speech being the hardest to understand and NV being the easiest. To approximately equate overall report score for each of the three manipulations, the modulation depths of the extracted amplitude envelopes were systematically altered. For NV speech, amplitude envelopes were square-root compressed (the amplitude of the envelopes extracted from each band was raised to the power of 1/2) and for sine-wave speech they were expanded by raising the amplitude of each band to the power of 1.25. For PT speech, the modulation depth of the extracted amplitude envelopes was not manipulated. Compressing the extracted amplitude envelopes reduces the modulation depth of the carrier signals in the vocoded output, rendering it harder to understand; expanding them increases the modulation depth and renders the output easier to understand (within the range used here), as it increases the amount of information from the envelope that is carried into the final distorted signal. Figure 3 shows the waveforms and spectrograms of each of the 3 types of vocoding applied to the same sentence. The vocoded sentences were then

[1] The difference between F0 and pulse repetition rate arises because the waveform of an alternating-phase complex consists of a series of pulsatile peaks, where the precise waveform shape of the odd-numbered peaks differs from that of the even-numbered peaks. When the harmonics of the complex are unresolved by the peripheral auditory system, as was the case here, the pitch of the complex corresponds with the pulse rate, which is equal to twice the F0 (Shackleton & Carlyon, 1994).

assembled into Distorted-Clear-Distorted (DCD) triplets, as in Experiment 1.

Procedure. The procedure was similar to that of Experiment 1. Participants were divided into nine groups of 12. There were three "No Transfer" groups, who received NV, SW, or PT speech throughout (4 blocks of 10 sentences), as DCD triplets, reporting what they could understand from each sentence after the first distorted presentation of each sentence, and subsequently hearing the CD versions as feedback. We also tested six "Transfer" groups (NV–PT, NV–SW, SW–NV, SW–PT, PT–NV, PT–SW) who heard 20 DCD sentence triplets vocoded using one carrier followed by a further 20 DCD sentence triplets vocoded using a different carrier. Once again we will refer to the first 20 sentences as "naïve" and the next 20 as "switch" in the transfer groups and "no-switch" in the no transfer groups. Written sentence responses following the initial presentation of each sentence were scored as for Experiment 1. As before, comparison of switch conditions with naïve and trained listeners can be used to establish the degree of generalization of learning among vocoding types.

Results

As in previous experiments, analyses were carried out on the data averaged over participants (the data for each participant were averaged within 10-sentence blocks). Mixed ANOVAs were conducted with test block as a within-subjects factor with two levels (first or second block of 10 sentences), and condition as a between-subjects factor with two levels (naïve vs. switch or no-switch vs. switch). A dummy variable for sentence group presentation order was entered as a between-subjects factor with four levels (AB, BA, CD, DC). Once again, effects of the block variable will not be reported. As in Experiment 1, comparisons between naïve and no-switch performance were carried out within participants and included only the data from 12 participants in the no-switch conditions. Other comparisons with naïve performance were between-participant and included all 36 data-sets from listeners in the naïve conditions.

SW. Results are shown in Figure 4a. Performance in the no-switch condition was significantly better than naïve performance, $F(1,8) = 141.074$, $p < .001$, partial $\eta^2 = 0.946$, indicating that learning occurred. Naïve SW performance differed from performance in the SW after NV group (though this was only marginally significant, $F_{(1,40)}=2.802$, $p = .051$, partial $\eta^2=0.065$) and differed from performance after training with PT, $F(1,40) = 2.907$, $p = .048$, partial $\eta^2 = 0.068$. Performance in the no-switch group was significantly better than in the two switch groups (from NV: $F(1,16) = 33.738$, $p < .001$, partial $\eta^2 = 0.678$; from PT: $F(1,16) = 17.541$, $p < .001$, partial $\eta^2 = 0.523$). These results show that performance on SW after training with PT and NV speech is intermediate between naïve and no-switch, suggesting a partial generalization of learning. Participants trained on vocoded speech with a different carrier signal were better than naïve listeners when tested with a SW carrier, but worse than listeners who had received prior training with that carrier signal.

PT. Results are shown in Figure 4b. Performance in the no-switch condition was significantly better than naïve performance, $F(1,8) = 134.439$, $p < .001$, partial $\eta^2 = 0.944$, indicating that learning occurred. Performance in the two switch conditions was significantly better than naïve PT performance (from NV: $F(1,40) = 5.749$, $p = .011$, partial $\eta^2 = 0.126$; from SW: $F(1,40) = 5.624$, $p = .011$, partial $\eta^2 = 0.123$). Performance in the switch conditions was significantly worse than that of the no-switch group (switch from NV: $F(1,16) = 3.101$, $p = .049$, partial $\eta^2 = 0.162$; switch from SW: $F(1,16) = 3.464$, $p = .041$, partial $\eta^2 = 0.178$). These results show that performance in the switch conditions is intermediate between the naïve and no-switch, suggesting partial generalization of learning.

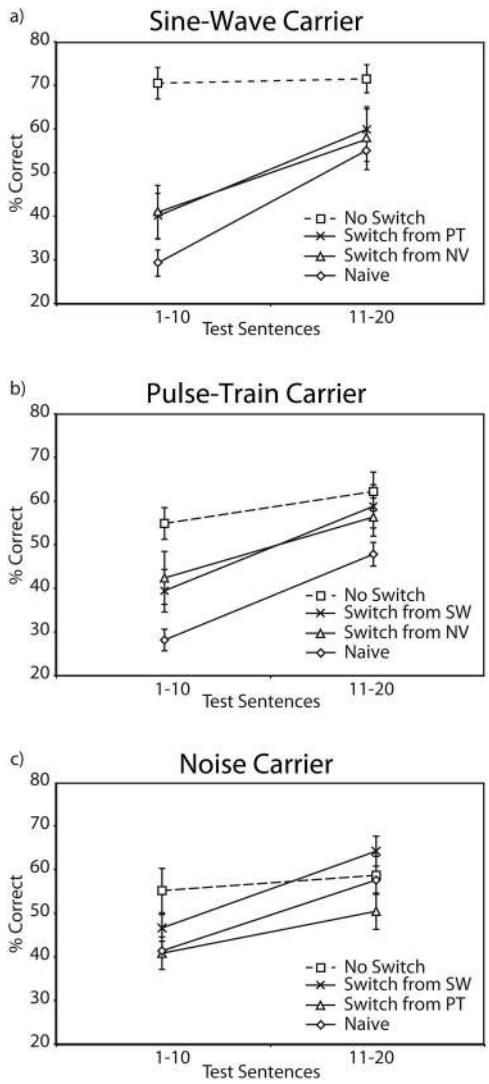

Figure 4. Results of Experiment 2, showing performance averaged over participants over 10-sentence blocks. Test sentences 1–10 and 11–20 indicate the first and second 10-sentence blocks in each condition. Error bars represent ± 1 *SEM*. (a) sine-wave vocoded speech, (b) pulse-train vocoded speech, (c) noise-vocoded speech. Naïve, $N = 36$; no-switch and switch, $N = 12$.

NV. Results are shown in Figure 4c. Performance in the no-switch condition was significantly better than naïve performance, $F(1,8) = 21.609$, $p < .001$, partial $\eta^2 = 0.730$, indicating that learning occurred. Naïve NV performance was somewhat worse than performance after SW though statistical significance was marginal, $F(1,40) = 2.124$, $p = .076$, partial $\eta^2 = 0.050$).

Performance on NV after PT was not significantly different from Naïve NV, although it was numerically lower, $F(1,40) = 0.957$, $p = .167$, partial $\eta^2 = 0.023$). Performance in the no-switch group was significantly better than in the NV after PT switch group, $F(1,16) = 5.559$, $p = .016$, partial $\eta^2 = 0.258$. However, no-switch performance did not differ significantly from performance in the NV after SW group, $F(1,16) = 0.084$, $p = .388$, partial $\eta^2 = 0.005$, despite the performance of the NV after SW group being only marginally better than naïve performance. These results might suggest that training with PT does not show significant generalization to NV, but that training with SW may generalize to some extent. Direct comparison of the NV after SW, and NV after PT performance shows that performance on NV is significantly better after SW than after PT, $F(1,16) = 5.653$, $p = .030$, partial $\eta^2 = 0.254$, further suggesting that training with PT does not generalize to NV while training with SW, at least partially, does.

Discussion

The results of this experiment show that transfer of perceptual learning of vocoded speech from one carrier signal is not complete. In nearly all cases, performance on switch conditions is significantly better than naïve, but significantly worse than no-switch. The only exceptions to this are NV after SW which is not significantly better than naïve but is not significantly different from no-switch performance, and NV after PT, which is not significantly different from naïve NV. This suggests that perceptual learning of vocoded speech partially transfers between vocoders that use different carrier signals. In Experiment 1, the slowly changing amplitude envelopes which convey speech information were at completely different frequencies in the two forms of vocoding tested. Yet, perceptual learning generalized significantly and completely between frequency ranges. In contrast, in Experiment 2, the same frequency bands and amplitude envelopes (differing only in modulation depth) were used with three different carrier signals, however complete generalization was no longer observed. This suggests that it is not enough for the amplitude envelope in specific frequency ranges to be similar in order to observe complete generalization; these amplitude envelopes must be imposed on the same fine-structure for complete generalization to be observed.

There are some differences in the ceiling levels of performance achieved by listeners in the no-switch conditions. Post hoc pairwise comparison of the mean scores in each of the no-switch conditions shows that performance is significantly better in the no-switch SW than no-switch NV condition ($p = .038$). Although there are no other significant pairwise differences (no-switch NV vs. no-switch PT, $p = .936$, no-switch PT vs. no-switch SW, $p = .120$), it could be said that ceiling performance on PT stimuli is intermediate between SW and NV. This trend may be because of the manipulations of the amplitude envelopes—envelope compression was applied to the NV stimuli, the envelopes of the PT stimuli were not altered and those of the SW stimuli were expanded. Because envelope compression renders stimuli less intelligible and expansion renders them more so, this trend is consistent with the manipulations imposed on the stimuli.

One intriguing finding is that there is some evidence for generalization from training on SW to testing on NV speech, although there is no compelling evidence for generalization in the opposite direction. This may be simply because of the relatively lower level of ceiling performance for the NV speech that introduces an apparent asymmetry. Given the relatively small overall improvement in performance for NV speech for the no-switch condition, it may in fact be that learning generalizes as effectively from SW to NV as the reverse. Indeed, a pairwise comparison of the mean improvement produced by SW training on NV report and that produced by NV training on SW report shows no significant difference ($p > .9$). The apparent asymmetry in generalization appears, therefore, to have more to do with the smaller improvement between naïve and no-switch performance used as reference conditions for generalization to NV speech rather than with an asymmetry in the directionality of generalization.

The apparent absence of improvement in NV report scores after training with PT speech is difficult to reconcile with the improvements observed in all the other switch conditions. We propose the following explanation. Rosenblith, Miller, Egan, Hirsh, and Thomas (1947) reported that listeners who had been exposed to pulse trains at rates of between 30 and 200 pulses per second reported that sounds presented immediately afterwards had acquired a peculiarly "metallic" timbre. This phenomenon has recently been examined in more detail by Gutschalk, Micheyl, and Oxenham (2008), who replicated the effect. Gutschalk et al.'s investigation showed that listeners' amplitude modulation detection thresholds were significantly elevated immediately after exposure to 100 pps sine-phase 100 Hz F0 harmonic pulse trains. Though the duration of the effect appears to be quite short, lasting in the region of 30 s, after 60 s of exposure to the pulse train stimulus, it is not clear from the existing data whether the effect of longer periods of exposure to pulse-trains could induce a longer-lasting aftereffect. Gutschalk and colleagues further showed that the effect persisted even in listeners who had had 16 h of practice on the task. Given that the speech information in NV speech is carried by amplitude modulation, it seems possible that elevated amplitude modulation detection thresholds would be deleterious to speech intelligibility. Although this account is highly speculative, it suggests a mechanism that might have imposed significant additional difficulty in using the information available in the NV speech for listeners previously exposed to PT stimuli.

General Discussion

The results of these experiments demonstrate that the perceptual learning of vocoded speech shows complete generalization across frequency region, but only incompletely generalizes across all the different carrier signals used in the investigation. In combination with existing results demonstrating that perceptual learning of vocoded speech generalizes to nontrained words (Davis et al., 2005; Hervais-Adelman et al., 2008) these findings help constrain the level of perceptual processing that is modified during learning. In developing a more detailed account of the perceptual learning of vocoded speech, we will discuss the anatomical implications of the two forms of auditory generalization assessed in the present experiments as well as some more general functional implications for accounts of speech perception and perceptual learning.

Generalization Across Frequency Region

Perceptual learning of NV speech generalizes from one frequency region to another, nonoverlapping, frequency range. This

result suggests that perceptual learning alters processes that apply equally across multiple frequencies, and are thus probably not occurring at levels of the auditory system which exhibit relatively narrow frequency tuning (i.e., all levels up to and including primary auditory cortex). Given the generalization between frequency regions that we have observed, we suggest that the changes underlying improved perception of NV speech are likely to modify representations beyond primary auditory cortex. This conclusion appears to be consistent with the results of a number of functional imaging studies in which differences between responses to intelligible and unintelligible vocoded speech are primarily observed in brain regions outside of primary auditory cortex, including anterior superior and middle temporal-lobe regions, and areas within the superior temporal sulcus (e.g., Davis & Johnsrude, 2003; Giraud et al., 2004; Obleser et al., 2006; Scott, Blank, Rosen, & Wise, 2000; Scott & Wise, 2004). Anatomical studies in humans and macaques indicate that multiple anatomically distinguishable regions intervene between primary auditory cortical "core" and the dorsal bank of the superior temporal sulcus (Fullerton & Pandya, 2007; see Hackett & Kaas, 2004, for a review of work in nonhuman primates). The cortex of the middle temporal gyrus is even further removed from auditory core cortex. As a consequence, it seems likely that the regions that are found to be sensitive to intelligible vocoded speech are at least two, and probably more, processing stages removed from primary auditory cortex.

One study that compared neural responses with vocoded speech before and after training (Giraud et al., 2004) observed differential neural responses in anterior superior temporal sulcus, and the middle and inferior temporal gyri, clearly well beyond primary regions. However, further studies which track changes in cortical responses during learning will be required in order to go further, and directly localize the neural systems that contribute to perceptual learning.

Our report of generalization of perceptual learning between frequency regions appears at odds with the results of Fu and Galvin (2003) who found that improved comprehension of frequency-shifted vocoded speech was specific to test stimuli with the same frequency shift. We can reconcile these apparently conflicting findings by recalling that Fu and Galvin manipulated not only the frequencies at which vocoded speech was presented, but also the degree of frequency mismatch between the original and the vocoded speech. Research on Helium speech (Belcher & Hatlestad, 1983; Morrow, 1971), indicates that shifting formant frequencies creates an additional source of difficulty for speech perception even without the loss of fine structure introduced by vocoding. In our experiments, there was a direct correspondence between frequencies contained in vocoded and clear speech; in other words, the amplitude envelopes that we extracted from source bands were always applied to carriers in the same frequency region as the source band. The results of Fu and Galvin may best be explained by suggesting that what fails to generalize over frequency regions is not perceptual learning of vocoded speech per se, but the learning of specific envelope-frequency band pairings.

Generalization Across Carrier Signal

In contrast to generalization across frequency region, our results suggest that perceptual learning of vocoded speech is somewhat specific to the carrier signal used during training. The three forms of vocoder used in Experiment 2 supply essentially the same amplitude envelope in the same frequency ranges and differ from each other primarily in the temporal fine structure that carries this envelope. Because the fine-structure itself probably does not provide any useful information for understanding the vocoded speech (in NV speech, for example, there is *no* fine-structure information available at all), it may be that the absence of generalization is attributable to listeners not learning to ignore the fine-structure information. If they were able to preferentially process the available envelope information without interference from the fine-structure, generalization between carriers would be more likely to occur.

This finding adds to a body of existing data which suggests that despite envelope information being sufficient for accurate perception, temporal fine-structure nonetheless makes an important contribution to speech perception. For instance, Smith et al. (2002), showed that temporal fine-structure of speech-noise chimeras contains speech content information, when those chimeras are processed with few frequency bands. Lorenzi and colleagues (Lorenzi et al., 2006) presented similar speech-noise chimeras to normally-hearing and hearing-impaired listeners, and showed that the latter were significantly impaired in comprehension of speech based on temporal fine-structure, although their perception of speech based primarily on amplitude envelope information was near normal.

A Functional Account of the Data

A functional description of the results is that listeners are learning to use the amplitude modulation information contained in the NV speech to perceive speech. They may therefore be learning to attend specifically to this information, no matter what frequency range it is carried in. However, the results from Experiment 2 suggest that although some general property of vocoded speech may be learned and then used to understand it, irrespective of carrier, something is learned that is specific to the carrier. Taken together this suggests that there may be distracting influences of the different carriers, which prevent listeners from attending specifically to the amplitude modulation information contained in the signals. In more general terms, considering the "source-filter" model of speech production (Dudley, 1939; Stevens, 1999), using different carriers in vocoders is equivalent to speech produced by different sources (i.e., different vocal chords). Given that in a natural situation the only way for speech to be produced by different sources is for it to be produced by different individuals, it seems reasonable for the speech perception system to fail to generalize learning between vocoders with different carriers. We therefore speculate that the impact of changes to the carrier signal reflects perception of this cue as a signal of two different talkers.

In contrast, changing the frequency range in which vocoded speech is presented has more in common with changes to the fidelity of the communication channel. This is perhaps analogous to the high-pass filtering that occurs when we hear speech over a radio with a very small speaker, or the low-pass filtering of speech heard through a door. In both cases, perception can be challenged, yet this does not fundamentally alter our perception of the identity of the speaker. Information acquired about speech presented through one of these filters might therefore be appropriately generalized to speech presented through another. In the next section we will consider evidence for perceptual learning of other forms of

variant speech that is similarly talker-specific, as well as addressing the implications of our findings for accounts of speech perception.

Implications for Accounts of Speech Perception

The two studies presented here provide evidence that perceptual learning of vocoded speech shows complete acoustic generalization in some cases (over frequency range) but not others (carrier signal). These results combine with phoneme categorization data in illustrating how perceptual learning of speech can show varying degrees of acoustic specificity. For instance, Eisner and McQueen (2005) found that perceptual learning of an ambiguous fricative is specific to the voice on which listeners were trained. In contrast, Kraljic and Samuel (2006) found that learning of a stop phoneme with an ambiguous voice-onset time (VOT) generalized both to a novel talker and to a different stop consonant. Such contradictory results seem difficult to reconcile with a simple account in which perceptual learning arises from a single level of representation which encodes a specific degree of abstraction from the acoustic input.

One way of reconciling these disparate results is to recall Goldstone's (1998) definition that the goal of perceptual learning is to "improve its [an organism's] ability to respond to its environment." In such an account the optimal degree of acoustic generalization will depend not just on acoustic changes experienced in the context of an experiment, but also on long-term knowledge of which sources of acoustic variation are potentially and reliably informative. So, for phoneme category learning studies, it might be that the acoustic form of fricatives is particularly idiosyncratic and encodes more speaker-specific information than stop-consonants. In such a situation, an optimal perceptual strategy would be to assume that information acquired from fricatives is speaker-specific and would not apply to a novel speaker. In contrast, if VOT is a more consistent cue then it could be advantageous for perceptual learning (of stop consonants) to generalize among speakers.

By this account, then, seemingly contradictory results from phoneme category learning could be explained by an account of perceptual processing of speech in which generalization of perceptual learning depends on the expected degree of between-speaker variability for different speech sounds. This speaker-specific learning scheme is consistent with findings from Nygaard and Pisoni (1998) and Nygaard, Sommers, and Pisoni (1994), that show talker-specific adaptation in speech perception (for instance, as evidenced by enhanced comprehension of speech in noise for familiar talkers). Although vocoded stimuli may not provide sufficient information for talker identification, the three different carrier types produce very different-sounding stimuli, which can be easily distinguished from one-another, and could be considered to be superficially more different from one-another than different talkers are.

Models of speech perception do not always explicitly address the issue of how variability in speech (such as between-talker variation) is handled. One class of models, the "Abstractionist" models, such as TRACE (McClelland & Elman, 1986), MERGE (Norris, McQueen, & Cutler, 2000), and the distributed cohort model (Gaskell & Marslen-Wilson, 1997) postulate a talker-normalization stage outside the scope of the model which strips away indexical information from speech to allow mapping of the processed input onto invariant representations. In such models, indexical information plays no role in perception. Another class of models, the "Episodic" models (e.g., Goldinger, 1996, 1998; Goldinger, Kleider, & Shelley, 1999), suggest that idiosyncratic properties of individual talkers' speech are stored in lexical memory and subsequently mediate word recognition (e.g. Goldinger, 1998; Goldinger et al., 1999). Talker-specific learning effects are consistent with episodic models that allow for indexical information to influence perception. However, these models must be modified to include some degree of pre-lexical abstraction to ensure that perceptual learning can generalize between different lexical items (see Hervais-Adelman et al., 2008; McQueen et al., 2006). Conversely, talker-specific learning effects are only compatible with abstractionist accounts if top-down perceptual learning processes are incorporated that modify pre-lexical representations following recognition of speech from specific talkers. Hence we can propose neither purely abstractionist nor episodic accounts, but rather a hybrid approach in which abstract representations are modified on the basis of talker-specific information. If the differences between types of speech vocoded using different carriers are considered analogous to between-talker differences, then the lack of convincing generalization between different types of vocoding seems to be more consistent with an account of speech perception such as this hybrid account in which indexical information plays a role in determining the extent of generalization.

A recent study by Dahan and Mead (2010) examines generalization of perceptual learning of NV phonemes at different syllable positions. They find that perceptual learning does not readily generalize between onset and coda of words, indicating that learning takes place at an acoustically-specific level that retains some coarticulatory information, rather than at a purely phonological level that has discarded such information. This lends further support to the notion that generalization of perceptual learning of vocoded speech is limited by acoustical similarity between trained and untrained stimuli.

An alternative interpretation that might more readily apply to vocoded speech would be to suggest that high- and low-pass filtering (Experiment 1) provides information about the acoustic properties of the communication channel rather than the speaker. Low-pass filtered speech sounds like it is heard through a closed door, whereas one might hear high-pass filtered speech through a tinny loudspeaker or a telephone. In such a situation, an optimal perceptual strategy would be to assume that the speaker remains the same (and hence, generalization would be expected). In contrast changing the carrier signal and hence temporal fine structure (Experiment 2) provides a more dramatic perceptual change which is heard as a new talker and hence the perceptual system is more resistant to generalization. While this account of our results is rather impressionistic, it at least suggests that further explorations of cross-speaker generalization in perceptual learning (both for clear and vocoded speech) would be productive.

Conclusion

These experiments show that the locus of the plasticity underlying perceptual learning of vocoded speech occurs at a frequency-general level, which is most likely found somewhere beyond primary auditory cortex. Learning to understand vocoded speech

shows complete generalization over frequency region, but is somewhat specific to the type of carrier signal used in the vocoder, indicating that perceptual learning of vocoded speech involves modifications to acoustic processes that are sensitive to temporal fine-structure. The specificity of learning to the surface acoustic form of the manipulation may be related to the talker-specificity that has previously been observed in perceptual learning of speech. Such specificity supports models of speech perception that take talker-specific information into account when adjusting perceptual processing. However, they correspond with neither a simple episodic model in which all idiosyncratic, acoustic properties of speech are stored, nor to a purely abstractionist model in which all talker-specific information is discarded during perception. Prelexical abstraction in speech is far from being a trivial problem (as extensively reviewed by Obleser & Eisner, 2009), and the challenge for future accounts of speech perception in both the episodic and abstractionist traditions is to explain which acoustic features of speech contribute to perception, and how they are encoded. Explorations of acoustic generalization in perceptual learning look set to play a role in developing these more complex accounts of speech perception.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Science, 8,* 457–464.

Altmann, G., & Young, D. (1993, September). *Factors affecting adaptation to time-compressed speech.* Paper presented at the Eurospeech 9, Berlin, Germany.

ANSI. (1997). *ANSI S3.5 1997: American national standard methods for calculation of the speech intelligibility index.* New York: American National Standards Institute.

Belcher, E. O., & Hatlestad, S. (1983). Formant frequencies, bandwidths, and Qs in helium speech. *Journal of the Acoustical Society of America, 74,* 428–432.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116,* 3647–3658.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance, 36*(3), 704–728.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23,* 3423–3431.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., 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,* 222–241.

Deeks, J. M., & Carlyon, R. P. (2004). Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent sound segregation. *Journal of the Acoustical Society of America, 115,* 1736–1746.

Demany, L., & Semal, C. (2002). Learning to perceive pitch differences. *Journal of the Acoustical Society of America, 111,* 1377–1388.

Dudley, H. (1939). The automatic synthesis of speech. *Proceedings of the National Academy of Sciences of the United States of America, 25,* 377–383.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics, 67,* 224–238.

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 108,* 1877–1887.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425,* 614–616.

Fitzgerald, M. B., & Wright, B. A. (2000). Specificity of learning for the discrimination of sinusoidal-amplitude-modulation rate. *The Journal of the Acoustical Society of America, 107*(5), 2916. ASA.

Fu, Q. J., & Galvin, J. J., III. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *Journal of the Acoustical Society of America, 113,* 1065–1072.

Fullerton, B. C., & Pandya, D. N. (2007). Architectonic analysis of the auditory-related areas of the superior temporal region in human brain. *Journal of Comparative Neurology, 504,* 470–498.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., & Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14,* 247–255.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 22,* 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory and Cognition, 27,* 328–338.

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

Greenwood, D. D. (1990). A cochlear frequency-position function for several species–29 years later. *Journal of the Acoustical Society of America, 87,* 2592–2605.

Grimault, N., Micheyl, C., Carlyon, R. P., Bacon, S. P., & Collet, L. (2003). Learning in discrimination of frequency or modulation rate: Generalization to fundamental frequency discrimination. *Hearing Research, 184,* 41–50.

Gutschalk, A., Micheyl, C., & Oxenham, A. J. (2008). The pulse-train auditory aftereffect and the perception of rapid amplitude modulations. *Journal of the Acoustical Society of America, 123,* 935–945.

Hackett, T. A., & Kaas, J. H. (2004). Auditory cortex in primates: Functional subdivisions and processing streams. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences III* (pp. 215–232). Cambridge, MA: MIT Press.

Hervais-Adelman, A., 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,* 460–474.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791–804.

Julesz, B. (1971). *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 11793–11799.

Kraljic, T., & Samuel, A. G. (2006). How general is perceptual learning for speech? *Psychonomic Bulletin & Review, 13,* 262–268.

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America, 106,* 2097–2103.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to

use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 18866–18869.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Science, 10,* 533.

Morrow, C. T. (1971). Speech in deep-submergence atmospheres. *Journal of the Acoustical Society of America, 50,* 715–728.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299–325; discussion 325–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–238.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60,* 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker contingent process. *Psychological Science, 5,* 42–46.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., . . . Rauschecker, J. P. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping, 27,* 562–571.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences, 13*(1), 14–19.

Peterson, G., & Barney, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21,* 785–794.

Raaijmakers, J. G. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology, 57,* 141–151.

Raaijmakers, J. G., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the Language-as-Fixed-Effect Fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41,* 416–426.

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 11800–11806.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129–156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: FMRI studies of semantic ambiguity. *Cerebral Cortex, 15,* 1261–1269.

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.

Rosenblith, W. A., Miller, G. A., Egan, J. P., Hirsh, I. J., & Thomas, G. J. (1947). An Auditory Afterimage? *Science, 106,* 333–335.

Roth, D. A.-E., Amir, O., Alaluf, L., Buchsenspanner, S., & Kishon-Rabin, L. (2003). The effect of training on frequency discrimination: Generalization to untrained frequencies and to the untrained ear. *Journal of Basic and Clinical Physiology and Pharmacology, 14,* 137–150.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Scott, S. K., & Wise, R. J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition, 92,* 13–45.

Shackleton, T. M., & Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *Journal of the Acoustical Society of America, 95,* 3529–3540.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303–304.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America, 118,* 3177–3186.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416,* 87–90.

Stevens, K. N. (1999). *Acoustic phonetics.* Cambridge, MA: MIT Press.

Appendix

List of Sentences Used in Experiments 1 and 2

A(1)	his wig fell on the floor
A(2)	the child left all of his lunch at home
A(3)	the student tried to move the desk
A(4)	his face showed that his team had lost the game
A(5)	the couple had been together for three years
A(6)	she loved stories about “fairies,” wizards and dragons
A(7)	the cattle were kept in the barn
A(8)	the fireman climbed down into the bottom of the tunnel
A(9)	he always read a book before going to bed
A(10)	the group of friends got a taxi home after they left the nightclub
B(1)	there were mice in the cave
B(2)	the noise was very loud and difficult to ignore
B(3)	they thought that the house was haunted
B(4)	He broke his leg when he fell off the horse
B(5)	he ironed his shirt before he wore it
B(6)	there was lettuce and cucumber in the salad
B(7)	the man read the newspaper at lunchtime
B(8)	the view from the top of the ridge was amazing
B(9)	the audience was quiet once the song had started
B(10)	it was the women that complained when the old bingo hall was closed
C(1)	there were books in the cellar
C(2)	it was too cold to go camping in the winter
C(3)	the thief started to sprint very fast
C(4)	the soup was kept in a carton in the fridge
C(5)	the building had a nest in its roof
C(6)	there was a really beautiful sunset that evening
C(7)	the whole sky was full of birds
C(8)	an angry crowd was turned back at the government building
C(9)	the carpet and the curtains were the same colour
C(10)	the children were hoping to play some hockey and rugby at their school
D(1)	the car drove over the cliff
D(2)	he left school before he had done his exams
D(3)	the coin was thrown onto the floor
D(4)	the sketch showed that the road would pass the school
D(5)	the bruise on his knee was quite painful
D(6)	the boy was able to conceal his cigarette
D(7)	her new skirt was made of denim
D(8)	the gambler lost most of his money at the races
D(9)	it is common for people to avoid the dentist
D(10)	the dessert was put in the oven at the start of the meal

Received February 15, 2008
Revision received March 17, 2010
Accepted April 9, 2010 ■