Journal of Experimental Psychology:
Human Perception and Performance
2010, Vol. 36, No. 3, 704–728

0096-1523/10/$12.00 DOI: 10.1037/a0017449

Context-Conditioned Generalization in Adaptation to Distorted Speech

Delphine Dahan and Rebecca L. Mead
University of Pennsylvania

People were trained to decode noise-vocoded speech by hearing monosyllabic stimuli in distorted and unaltered forms. When later presented with different stimuli, listeners were able to successfully generalize their experience. However, generalization was modulated by the degree to which testing stimuli resembled training stimuli: Testing stimuli's consonants were easier to recognize when they had occurred in the same position at training, or flanked by the same vowel, than when they did not. Furthermore, greater generalization occurred when listeners had been trained on existing words than on nonsense strings. We propose that the process by which adult listeners learn to interpret distorted speech is akin to building phonological categories in one's native language, a process where categories and structure emerge from the words in the ambient language without completely abstracting from them.

Keywords: speech recognition, noise-vocoded speech, learning, generalization

Comprehending speech requires mapping the acoustic signal onto mental representations that correspond to linguistic units such as words or morphemes. Much of the complexity associated with this categorization process originates from the "lack of invariance" issue, famously demonstrated by Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967): The acoustic realization of phonological units such as phonemes varies greatly across speakers and across contexts. Because no naturally produced speech sound is acoustically identical to any of the sounds a listener has previously encountered, speech perception requires the generalization of past experience to novel instances.

Although people can usually understand the speech of any speaker of their dialect with no prior exposure to this speaker, there is ample evidence that listeners' familiarity with a talker facilitates processing of that talker's speech (e.g., Nygaard & Pisoni, 1998). This indicates that listeners are continuously learning from exposure to a talker's speech and adjusting their categorization scheme to improve the processing of this speech. A growing body of research has begun to investigate the nature of this adaptation (Clarke & Garrett, 2004; Kraljic & Samuel, 2006; Norris, McQueen, & Cutler, 2003). Building on this work, the current study examined the phonetic and lexical conditions under which listeners generalize their experience with spoken stimuli to novel ones.

A fruitful approach to examining listeners' adjustment to speech has been to expose listeners to synthesized or artificially distorted speech and to assess their improvement in interpreting this speech with increasing exposure. Different kinds of speech have been used, such as time-compressed speech (e.g., Dupoux & Green, 1997; Peelle & Wingfield, 2005), synthetic speech (e.g., Greenspan, Nusbaum, & Pisoni, 1988; Schwab, Nusbaum, & Pisoni, 1985), sine-wave analogues of speech (e.g., Remez, Rubin, Pisoni, & Carrell, 1981), or noise-vocoded speech (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Typically, participants are exposed to utterances and asked to report what they have understood. Improvement in word-identification rates, even for novel utterances, reveals generalization of learning. Feedback provided during training has been shown to speed learning (Davis et al., 2005).

Other work has exposed listeners to speech in which instances of a single sound category, such as the phoneme /s/, have been artificially altered to be ambiguous between two phonemic categories, e.g., /s/ and /f/ (e.g., Norris et al., 2003). The use of a single, well-specified phonetic alteration permits close examination of the change in listeners' phonetic categories. Learning in these studies is typically assessed by examining how listeners' categorization of the /s/- and /f/-like sounds changes. The basic result is that listeners use linguistic cues to identify the interpretation of the ambiguous sound that is most likely to be correct, and, having done so, explicitly label instances of the ambiguous sound as being members of that category (Eisner & McQueen, 2005; Kraljic & Samuel, 2006).

The use of novel forms of speech, whether considering only one phonemic category or the entire sound repertoire, ensures that listeners have no prior experience with the phonetic changes under study. As a result, the similarity structure of the training and test materials can be controlled, with the assurance that participants enter the experiment without specific knowledge of the accent or other distortion. If the procedure that converts ordinary speech into the distorted version is systematic or rule-based, then listeners'

Delphine Dahan and Rebecca L. Mead, Department of Psychology, University of Pennsylvania.

This work was supported by the National Science Foundation under Grant No. 0433567 and the National Institutes of Health (1 R01 HD 049742-1). We thank Matt Davis for kindly sharing with us the scripts used to generate the noise-vocoded version of our utterances, and Rachel Barreto for her help in conducting Experiment 1. We are also grateful to Saul Sternberg and the participants of the Cog Tea seminar for their helpful comments and suggestions on part of this work, and to Sarah Drucker for her comments on the manuscript. Experiments 1 ad 2 were presented at the 12th Annual Conference on Architectures and Mechanisms for Language Processing, Nijmegen, The Netherlands.

Correspondence concerning this article should be addressed to Delphine Dahan, Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Room 302C, Philadelphia, PA 19104. E-mail: dahan@psych.upenn.edu

implicit discovery of this systematicity can be evaluated by probing for generalizations that accord with the rule but that were not specifically exemplified in training. In this way, perceptual adaptation can be used as a tool for studying not only adaptation per se, but more general questions concerning the units over which perceptual categorization operates.

In the present study, we examined generalization in speech-sound identification over different phonetic environments and over different talkers. We exposed listeners to English utterances that had been distorted using a noise-vocoding procedure, a transformation (described in more detail below) that removes much of the spectral information present in speech while preserving temporal cues. The result is often described as sounding like a harsh whisper (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). This form of distortion has been previously used, most notably by Davis et al. (2005), to investigate perceptual learning and generalization. Because our study was partly motivated by Davis et al. (2005), it is useful to review their findings. Davis et al. exposed people to multi-word utterances in both original and distorted versions. After this training, listeners' ability to identify words in novel distorted sentences was tested. Trained listeners were able to correctly identify more words than untrained listeners, including some words that they had not heard during training. This led Davis and colleagues to conclude that the locus of the learning is sublexical: People learned to associate the acoustic patterns associated with subparts of the trained words to their respective categories, and were able to apply this knowledge to novel acoustic patterns if they were perceived as similar. A more direct test of generalization to untrained words has since been reported by Hervais-Adelman, Davis, Johnsrude, and Carlyon (2008), where participants were trained and tested on isolated words. Consistent with their prior conclusion, listeners were able to generalize their learning to novel, untrained words. Evidence for generalization after a short amount of training with a novel form of speech was also reported by Francis and Nusbaum (2000).

Building on this finding, we asked whether the categories over which listeners generalize their past experience with noise-vocoded utterances correspond to phonemic categories. Evidence that people can transfer their knowledge of the distortion acquired from some words to different words indicates that this generalization involves sublexical categories, that is, categories smaller than words. This is because if listeners learned only the distorted forms of individual words as wholes, without decomposing them into smaller elements, they would have no basis for generalizing this knowledge to additional words. However, the nature of these categories has not been investigated, including whether the categories correspond to phonemes. The concept of phonemes is central to capturing a language's phonology because the phonological rules that dictate the sound structure of words in the language are often best expressed in terms of phonemes, thus abstracting from the phonetic realization that the sounds take. As an illustration, consider the morphological rule of the gerund formation in English, in which the verb stem combines with the inflection *ing*. If words were solely represented by their overt, phonetic realization, the gerund *cutting* in American English would not comply well with the rule: The sounds depicted by the grapheme *t* at the end of the stem *cut* and within the gerund *cutting* are very different. If, however, the sounds of words are internally represented as members of phonemic categories, *cutting* and *seeing* conform equally well to the gerund-formation rule because the often unreleased consonant at the end of *cut* and the flap in *cutting* are described as two different realizations of the same underlying phonemic category /t/.

Thus, a language's phonological rules are best expressed if sounds that may be acoustically quite distinct are grouped into a common, functional category. However, whether such categories are involved at all during the perception of speech has been a debated question. Here, we examined whether phonemic categories are recruited when adapting to a novel form of speech. More specifically, when learning to process noise-vocoded speech, listeners may generalize what they have learned from one member of the phonemic category to the entire category. Such a generalization, however, would involve non-trivial computation. As pointed out earlier, the acoustic realization of the same phoneme varies greatly across contexts: The acoustic characteristics of a consonant can be dramatically affected by the vowel that precedes or follows it, syllable internally, and listeners are exquisitely sensitive to these coarticulatory dependencies (Liberman, Delattre, Cooper, & Gerstman, 1954; Warren & Marslen-Wilson, 1987; Whalen, 1991). The position of a consonant within a syllable (i.e., as the onset or the coda of the syllable) has also well-established effects on its realization (Brownman & Goldstein, 1995; Sproat & Fujimura, 1993), leading some to question whether allophonic variants of the consonants are represented as belonging to the same category (Bybee, 2001). There is also ample evidence that listeners can distinguish the allophonic realizations of the same phoneme as a function of its syllable position in order to hypothesize syllable and word boundaries in the speech stream (Christie, 1974; Lehiste, 1960; Nakatani & Dukes, 1977; Shatzman & McQueen, 2006; Spinelli, McQueen, & Cutler, 2003). Thus, generalization from exposure to one distorted sound to any sound within the same phonemic category would require the modeling of contextual influences on the acoustic realization of the distorted sound in order to learn how the noise-vocoded transformation renders a given phoneme once the influence of its context has been factored out.

If people were to generalize their experience with one member of a phonemic category to any of its members, this would provide compelling evidence that people can retrieve properties that are specifically associated with a phoneme despite contextual variations on its acoustic rendition. This result, in turn, would lead support to the view that the process applies to speech perception more generally. Some recent finding suggests such a process may be taking place: Francis, Nusbaum, and Fenn (2007) showed that people trained to understand synthetic speech have learned to shift weight given to acoustic correlates of phonetic contrasts. Alternatively, listeners' ability to transfer their learning may not apply equally well to any member of the phonemic category; instead, generalization may be affected by the similarity between the phonetic context in which the consonant was heard during training and that in which it occurred at testing. To the best of our knowledge, no study on speech adaptation has directly addressed this question. McQueen, Cutler, and Norris (2006) showed that listeners can extend their experience with an ambiguous sound at the end of [s]- or [f]-ending words to different [s]- or [f]-ending words or to a vowel-consonant syllable containing the same ambiguous sound. However, the ambiguous sound was acoustically the same at training and testing. Thus, while this result demonstrates that people were able to isolate the ambiguous sound from the words in which it appeared and apply the same interpretation across words, these results do not directly address the issue of contextual phonetic dependence in the generalization of category learning. In the present study, we asked if learning to recognize a member of a phonemic category in a distorted utterance generalizes to other members of the same phonemic category when aspects of the realization of these members are known to lawfully vary, either because of their phonetic contexts or because of cross-talker differences.

Cross-talker generalization has been examined in a number of studies involving atypical forms of speech (Bradlow & Bent, 2008; Dupoux & Green, 1997; Kouider & Dupoux, 2005). Learning to interpret artificially altered or foreign-accented speech appears to generalize quite readily from one talker to another, provided that the talkers have a similar native-language background or have had their speech altered in the same manner (but see Eisner & McQueen, 2005; Kraljic & Samuel, 2006, for reports of talker-specific adaptation when the altered sounds contain enough spectral information to be traced as originating from a given speaker). This suggests that the mapping that listeners learned operates on a speaker-independent representation of the sensory input, or only involves aspects of the acoustic signal that are stable across speakers. In the present study, we tested for evidence of both cross-talker and phonemic-category generalization to assess whether these two kinds of generalization can be observed under the same training conditions. There are good reasons to hypothesize that they differ. Generalization across talkers is a necessary component of any flexible speech-perception system. But, as reviewed above, whether the recognition of words in speech requires conversion of the speech signal into a series of phonemic categories has been questioned. For example, listeners might implicitly group instances of /t/ in only some environments as "the same," e.g. in *bat* and *cat*, without considering other members of the phoneme class as the same, e.g. in *tap* and *stop*. The fact that these sounds are all transcribed with "t" in English orthography (and that they are considered as instances of the phoneme /t/) does not imply that they are categorized as the same in language perception. Likewise, learning to adapt to distorted speech might not generalize equally to all instances within a phonemic category.

Although the question pertaining to generalization of learning was the main focus of the research presented here, we also examined the contribution of a constraint that Davis et al. (2005) reported as critical for learning to take place, namely training using words familiar to the participants. Norris, McQueen, Cutler, and colleagues (Eisner & McQueen, 2005; McQueen et al., 2006; Norris et al., 2003) showed that adjustment of phonetic-category boundaries could happen because of lexical information that helps listeners learn to interpret an ambiguous stimulus. In the basic paradigm they devised, participants were exposed to ambiguous tokens (mid-way between "s" and "f") occurring in the final position of s-ending or f-ending words. For one group of participants, only an "s" interpretation of the ambiguous sound formed an existing word, while for the other group, only an "f" interpretation of the ambiguous sound formed an existing word. Listeners' retuning of their phonemic category "s" or "f" was assessed by a subsequent phonemic categorization test: Listeners who had been exposed to the ambiguous s-f tokens in "s"-ending words were more likely to label an ambiguous s-f stimulus as an "s" than listeners who had been exposed to the same ambiguous s-f tokens in "f"-ending words. Because no such difference was found when listeners were first exposed to ambiguous s-f tokens embedded in strings for which neither an "s" or "f" interpretation formed a real word, the perceptual retuning was attributed to a feedback signal from the lexicon.

Davis et al. (2005) made the stronger claim that this lexical knowledge is *necessary*, and not merely beneficial, to learning, even in conditions where listeners receive explicit feedback about the phonological contents of the stimulus after each distorted utterance. In this study, listeners were trained to interpret noise-vocoded speech by hearing 20 8-to-13-word sentences. Each sentence was presented in its distorted form, then in its original undistorted form, then in its distorted form once again. Some participants heard sentences composed of real words, while others heard sentences composed of phonotactically legal nonwords. Following training, new distorted sentences were presented. These sentences were the same across groups and were composed of real words. Listeners were asked to identify as many words as they could in the sentence. Participants correctly reported more words than naïve participants (who had not received any training prior to testing) only when they had been trained on sentences containing real words. Davis et al. concluded that hearing distorted real words (as opposed to nonsense strings) was critical to learning to interpret the distorted form of speech.

Although intriguing, the Davis et al. results raise the question of why lexical knowledge was critical to learning to map acoustic patterns onto phonological categories when listeners were provided with the clear, unaltered version of the sentence after each distorted utterance. An explanation for Davis et al. (2005)'s results hinges on the hypothesis that the feedback was much less effective when the sentences were composed of nonwords because of short-term memory limitation. Recall that sentences were 8 to 13 words long. Listeners may not have been successful in holding such long nonsense strings in memory while attempting to interpret the noise-vocoded signal. Davis and colleagues have since published data suggesting that training with mono- and bi-syllabic nonword stimuli, for which short-term memory demands are minimal, was just as effective as training with mono- and bi-syllabic word stimuli (Hervais-Adelman et al., 2008). However, their lexicality manipulation did not control for the phonetic similarity between the subcomponents of the training stimuli with those at testing: While they ensured that phonemes occurred (roughly) the same number of times in their word and nonword stimuli, they did not match these two types of stimuli on the distribution of these phonemes within the stimuli. Because, as the current study will show, generalization from training to testing is strongly influenced by the similarity of the phonetic context in which exemplars of a phonemic category were encountered at training (i.e., the position of a consonant within the word or syllable and the identity of the vowel that flanked that consonant), a test of lexicality on learning requires controlling for phonetic similarity across word and nonword stimuli. The experiments reported here did just that.

Aside from the difference between word and nonword training stimuli on their memory demands, training with word stimuli may be more effective because lexical knowledge constrains the set of possible phonemic strings, and this knowledge could contribute to the decoding of an inherently ambiguous acoustic signal. Following the seminal work by Miller, Heise, and Lichten (1951), which applied Shannon (1948)'s theory of communication to speech perception (cited in Miller, Bruner, & Postman, 1954), Boothroyd and Nittrouer (1988) showed that the identification of speech segments under noisy conditions reflects contextual influence: While the segments that compose a nonsense speech string are identified more or less independently from one another, this independence is mitigated when these segments form an existing word (see also Benkí, 2003). Because a word context restricts the set of possible segmental strings, segment identification is more accurate in words than in nonsense contexts. Here, we examined whether this principle applies to learning to interpret noise-vocoded stimuli when feedback, in the form of the clear versions of each utterance, is provided.

The present study reports the results of four experiments. Experiments 1 and 2 examined the impact of the similarity between the phonemic composition of training and testing stimuli on listeners' ability to generalize the knowledge acquired during training. This similarity was manipulated by assessing listeners' recognition of consonants that they had encountered at training when these consonants occurred in the same syllabic position and/or flanked by the same vowel as in training or not. Experiments 1 and 2 also compared listeners' ability to generalize when they had been exposed to word or nonword stimuli at training. Experiments 3 and 4 directly compared listeners' ability to generalize their experience with particular speech sounds when these sounds were produced in a different phonetic context and by a different speaker.

Experiment 1

Experiment 1 examined the impact of acoustic similarity and lexicality on listeners' ability to interpret noise-vocoded speech. All participants were tested on the perception and identification of the same set of noise-vocoded CVC (Consonant Vowel Consonant) words. However, the characteristics of the training regime they received prior to testing varied. The similarity between the training stimuli and the testing stimuli was varied, and the training stimuli were either existing English words or nonsense syllables. We manipulated the similarity between training and testing stimuli by varying, across participants, the phonetic overlap between training and testing stimuli. Although never identical to the training stimuli, testing stimuli could share with the training stimuli (1) their onset-nucleus or rime components, or (2) their onset or coda consonants only, or (3) their consonants but in different syllable positions. We also examined the role that hearing existing words during training might play in learning to interpret new noise-vocoded stimuli. We compared the performance in identifying distorted testing words by participants who had been exposed to noise-vocoded nonword stimuli with feedback to the performance of participants who had been trained with words and to the performance of participants who had had no pre-exposure with noise-vocoded speech.

To simplify the design and limit the number of conditions, the lexicality and similarity factors were not fully crossed; only one similarity condition (i.e., the consonant overlap) was used to compare performance between participants trained on words and those trained on nonwords.

Method

Participants. Seventy-five students from the University of Pennsylvania took part in this experiment in exchange for course credit. All were native speakers of English. They provided informed consent prior to the beginning of the experiment. The protocol was approved by the University of Pennsylvania Institutional Review Board.

Materials. The 19 English consonants that can legally appear at syllable onset and coda were divided into two sets, roughly equating the sets in terms of manner of articulation. Nine consonants were assigned to the "onset" set (/b/, /t/, /g/, /f/, /s/, /θ/, /tʃ/, /m/, and /l/), the remaining 10, to the "coda" set (/p/, /d/, /k/, /v/, /z/, /ð/, /dʃ/, n/, /r/, and /ʃ/). Each consonant in each set was used between 1 and 3 times in their respective position (i.e., as onset or coda) to form twenty CVC existing English words. Half of the words had a low frequency of occurrence in the language (with a frequency of less than 4 occurrences per million words, according to the CELEX word form database, Baayen, Piepenbrock, & Gulikers, 1995), and the other half, a much greater frequency (ranging from 17 occurrences to 941 occurrences per million words). These 20 CVC words were used as testing words. They are listed in Appendix A.

Each of the 20 test words was used to create 8 different training stimuli, two for each of the four training sets (see Table 1). Among the training sets, the training stimulus could resemble the test word (e.g., *chive*) in one of three ways: (1) by having the same onset and vowel, or the same vowel and coda, e.g., *chire* and *tive* (the onset-nucleus/rime overlap condition); (2) by having the same onset or the same coda, but a different vowel, e.g., *chaz* and *gav* (the onset/coda overlap condition)[1]; (3) by having the onset displaced to coda position and nothing else matching, or having the coda displaced to onset position and nothing else matching, e.g., *neech* and *vum* (the consonant overlap condition). The consonant overlap condition was implemented with both nonword training stimuli and word training stimuli, yielding the four training sets.

An additional constraint on the training stimuli pertained to the particular consonants in the training stimuli that did not match their "source" test word (e.g., the consonants *r* and *z* in *chire* and *chaz*). The selection of these consonants depended on the training group. In the onset-nucleus/rime overlap group and the onset/coda overlap group, all of the consonants appearing as codas in the test words also appeared as codas in the training words, and likewise the set of consonants appearing as onsets in the test words also appeared as onsets in the training words. In the consonant overlap groups, the coda consonants of the test words were used only as onsets in training, and the onset consonants of the test words were used only as codas in training. Please see Table 1 for an illustration of the stimulus types. The result of this careful assignment of sounds across training and testing conditions was that listeners had heard a given sound within a particular position or context at training or not, depending on the training group they were assigned to. The consequences of this experience were evaluated at test. In each of the training sets, each consonant occurred 4 times on

---

[1] The training set for the onset-coda overlap group mistakenly contained four items that overlapped with testing items by the onset consonant and the nucleus, rather than only by their onset consonants.

Table 1

Description of Four Training Stimuli Sets (With Examples)

Lexicality of training stimuli	Nonword training stimuli			Word training stimuli
Training-testing overlap	Onset-nucleus rime overlap	Onset coda overlap	Consonant overlap	Consonant overlap
Testing stimulus $C_aV_1C_b$ chive	$C_aV_1C_{\{C\}}$ chire $C_{\{O\}}V_1C_b$ tive	$C_aVC_{\{C\}}$ chaz $C_{\{O\}}VC_b$ gav	$C_{\{C\}}VC_a$ neech $C_bVC_{\{O\}}$ vum	$C_{\{C\}}VC_a$ ditch $C_bVC_{\{O\}}$ vase

Note. The subscripts a and b indicate consonant Identity; the subscripts {O} and {C} indicate that the consonants were selected from the onset or coda sets, respectively. The subscript 1 indicates that the same vowel as in the testing stimulus was used at training (see text for details).

average (varying between 3 and 6). Appendix A lists the complete stimuli.

In addition to the 20 critical testing words, 10 words were included to play the role of practice trials. Such trials seemed especially important for the group of naïve listeners who were not exposed to noise-vocoded utterances at training and who, we suspected, would be at first disconcerted by their initial encounter with noise-vocoded speech. With the exception of one word, the practice testing words contained consonants that were not present at training.

All training and testing stimuli were read by the second author, a female native speaker of American English, and recorded directly onto a computer (sampling at 22 kHz with 16-bit resolution). Each stimulus was edited and a noise-vocoded version of each utterance was generated using the speech editor and synthesizer *Praat* (Boersma & Weenink, 2005). The procedure, identical to that used by Davis et al. (2005), consists of defining 6 logarithmically spaced bands in the frequency domain between 50 and 8000 Hz (with cutoff frequency values of 229, 558, 1161, 2265, and 4290 Hz) and extracting the amplitude envelope from each band. The envelopes are then applied to band-limited white noise in the same frequency ranges. Combining these amplitude-modulated noise bands generates utterances where the temporal and amplitude cues present in the original utterance are preserved, but where the spectral detail within each band is lost (see Davis et al., 2005, page 225, for further detail). The average amplitude of all stimuli was subsequently normalized to 55 dB Sound Pressure Level.

Procedure. Participants were randomly assigned to one of the five training groups. Participants in the "onset-nucleus rime overlap" group heard training *nonword* stimuli with consonants occurring in the same position and embedded in the same onset-nucleus or rime as they would be in the testing stimuli. Participants in the "onset-coda overlap" group heard training *nonword* stimuli with consonants that occurred in the same syllable position as they would be in the testing stimuli, but with no onset-nucleus or rime overlap. Participants in the "consonant overlap" group heard training *nonword* stimuli with the same consonants as those present in testing stimuli, but in a different syllable position and with no vowel overlap. Participants in the "word" group heard training *word* stimuli with consonants that would occur in the testing stimuli but in a different syllable position and with no vowel overlap. Finally, participants in the "naïve" group heard the same training nonword stimuli as those in the "consonant overlap" group, but only in their undistorted version. Thus, participants in the naïve group lacked exposure to noise-vocoded speech prior to testing, although they did have some exposure with the talker's undistorted speech.

Participants were individually tested in a quiet room. Instructions described the study as an examination of how people learn to adapt to a novel form of speech, one in which natural utterances have been artificially distorted to sound like a harsh, noisy whisper. Depending on which training group they were assigned to, participants were instructed that they would hear a series of real or made-up words. Their task was simply to listen to each word very carefully. Each word was presented three times, first in its distorted form, then in its original, non-distorted version, and finally in its distorted form again. Participants who were assigned to the naïve group heard the original recordings of the consonant-overlap (nonword) training items three times successively. After the training phase, a series of isolated real English words, produced by the same speaker as the stimuli heard earlier, were presented in their distorted form once. Participants' task was to guess which real word they heard and to type their answer on the keyboard. Participants were encouraged to give their best guess, even if unsure, and to take as much time to answer as needed. Participants moved to the next distorted stimulus by pressing the space bar. The experiment took less than 20 min, including debriefing.

The order with which the training stimuli in each group were presented was varied by randomly dividing the 40 stimuli into four groups of ten stimuli and randomizing the presentation order of these groups across participants. During testing, the same random trial order was used for all five groups.

Results

Participants' responses were coded for the accuracy of each of the test word's consonants and the accuracy of the word as a whole. A consonant response was coded as accurate if it corresponded to that of the stimulus exactly. Responses that contained a consonant cluster, either as onset or as coda, were considered inaccurate, even when the cluster contained the stimulus consonant. (A more lenient coding, for which a response was coded as accurate if the first consonant of the onset consonant cluster or the last consonant of the coda consonant cluster corresponded to the stimulus onset or coda consonant, respectively, yielded comparable results.)

We analyzed the data using multilevel logistic regressions. There is a tradition of analyzing categorical outcomes, such as outcomes coded as correct or incorrect, using analyses of variances over proportions. This approach is problematic, however, as recently reviewed by Jaeger (2008). Indeed, proportions are bounded (i.e., they vary between 0 and 1), and the variances of two binomially distributed conditions are usually not homogeneous because their variance is directly affected by how distant from .50 the sample means are. Logistic regressions correct for these issues by transforming proportions into log-odds. A perhaps less familiar aspect of the analyses we present here concerns the use of hierarchical (multilevel) modeling. Ordinary logit models assume that observations are independent from one another. However, this assumption is violated when observations are clustered, as they necessarily are when they originated from the same participant or are responses given to the same item. In mixed logit models, variations associated with these clusters are modeled as random effects. This significant advantage of mixed regressions over ordinary linear regressions is contributing to its growing popularity within the linguistic and psycholinguistic communities (see Jaeger, 2008; Baayen, 2008).

Consonant identification. Figure 1A presents the proportion of correct consonant identifications at test as a function of participants' training groups.[2]

First, we examined the impact of the similarity between the consonants heard at training and testing on participants' ability to correctly identify them. As apparent on the figure, greater overlap between training and testing stimuli resulted in better performance: Participants in the onset-nucleus rime overlap condition correctly identified the consonants in testing words more frequently than participants in the onset-coda overlap condition, and the latter performed better than those in the consonant overlap condition.

To statistically evaluate these differences, we conducted two logistic regressions on consonant identification accuracy. For each analysis, following Jaeger (2008) and Baayen (2008), we fitted a number of mixed logit models to the data using the function lmer (from the lme4 library) from the R statistical package (version 2.7.0, see Jaeger, 2008; Baayen, 2008, for more details). These models differed in terms of their random effects only, and the model that fitted the data best with the fewest degrees of freedom is reported.

Syllable-position match effect. The first analysis contrasted participants' performance when the consonants at testing had been presented in the same position as at training and in the opposite position (i.e., the contrast between the consonant overlap group and both the onset-coda overlap and the onset-nucleus rime overlap groups). For fixed effects, our model included the main effects of the position of the tested consonant (onset or coda), its match with the position the consonant occupied at training (same or different syllable position), the lexical (log) frequency of the testing word, as well as the interaction between the consonant's position at testing and its position match. In terms of random effects, we established that the simplest and best fitting model was the model that estimated different intercepts (i.e., accuracy level) for each participant, each item, and each consonant. In this model, there was a main effect of the syllable-position match between training and testing, with a coefficient of 0.72 ($p = .01$). This indicates that the odds of correctly identifying a testing consonant presented in the same position as in training are 2 times greater than the odds of correctly identifying a testing consonant that was heard in a different position at training. No other effect contributed significantly to predicting consonant identification.

Vocalic context effect. The contribution of the vocalic context in which the tested consonants were heard was evaluated by comparing performance of the onset-nucleus rime overlap group with that of the onset-coda overlap group. For both groups, consonants at testing occurred in the same position as at training. However, these consonants were systematically flanked by the same vowel in the onset-nucleus rime overlap group only. To test whether the vowel match had a systematic contribution to consonant identification, we fitted a mixed logit model that included, as fixed effects, the vowel match, the position of the consonant, and the frequency of the testing word. None of these factors were found to significantly predict consonant identification.

We examined whether the lack of a significant difference between the onset-nucleus rime overlap and the onset-coda overlap conditions might have resulted from the presence of four items in the onset-coda overlap training set that shared their vowel with testing items, in effect making the onset-coda overlap group training more like that of the onset-nucleus-rime overlap group. We fitted a mixed logit model to a subset of the data where the four testing items had been removed. In this model, the vowel match had a marginal contribution to predicting consonant identification, with a coefficient of 0.4 ($p = .09$). This indicates that the odds of correctly identifying a consonant in the testing word was 1.5 times greater when the consonant had been heard at training flanked with the same vowel as at testing than when it had occurred in a different vocalic environment.

Effect of prior exposure with noise-vocoded speech. Experiment 1 also examined whether training with nonsense syllables (nonwords) led to better consonant identification than no training at all. We compared performance of listeners who received nonword training (i.e., participants in the onset-nucleus rime overlap group, in the onset-coda overlap group, and in the consonant overlap group) with performance of listeners who received no noise-vocoded training (i.e., the naïve group). As apparent in Figure 1A, the performance of participants who heard distorted nonword stimuli (i.e., the onset-nucleus rime overlap group, the onset-coda overlap group, and the consonant overlap group) correctly identified test consonants more frequently that did participants in the naïve group (although the size of the difference varied as a function of the similarity between training and testing stimuli). To confirm this, we fitted a mixed logit model to our data that included, as fixed effects, the training status (nonword training or no training), the position of the test consonant, and the lexical frequency of the tested word. Here again, the simplest model that best fitted the data was one that estimated different intercepts for each participant, item, and consonant (random effects). In this model, training status had a significant contribution to the probability of correctly identifying consonants at test, with a coefficient of 0.84 ($p = .00015$). The value of the coefficient indicates that the odds of correctly identifying a testing consonant presented are 2.3 times greater when people were trained with nonword stimuli than when people received no noise-vocoded speech training. No other effect contributed significantly to predicting consonant identification.

---

[2] On Figure 1 and all subsequent figures, the data are expressed in proportion of correct consonant identification computed across participants, with standard errors capturing inter-participant variability. However, for the reasons expressed above, analyses are conducted on logit.

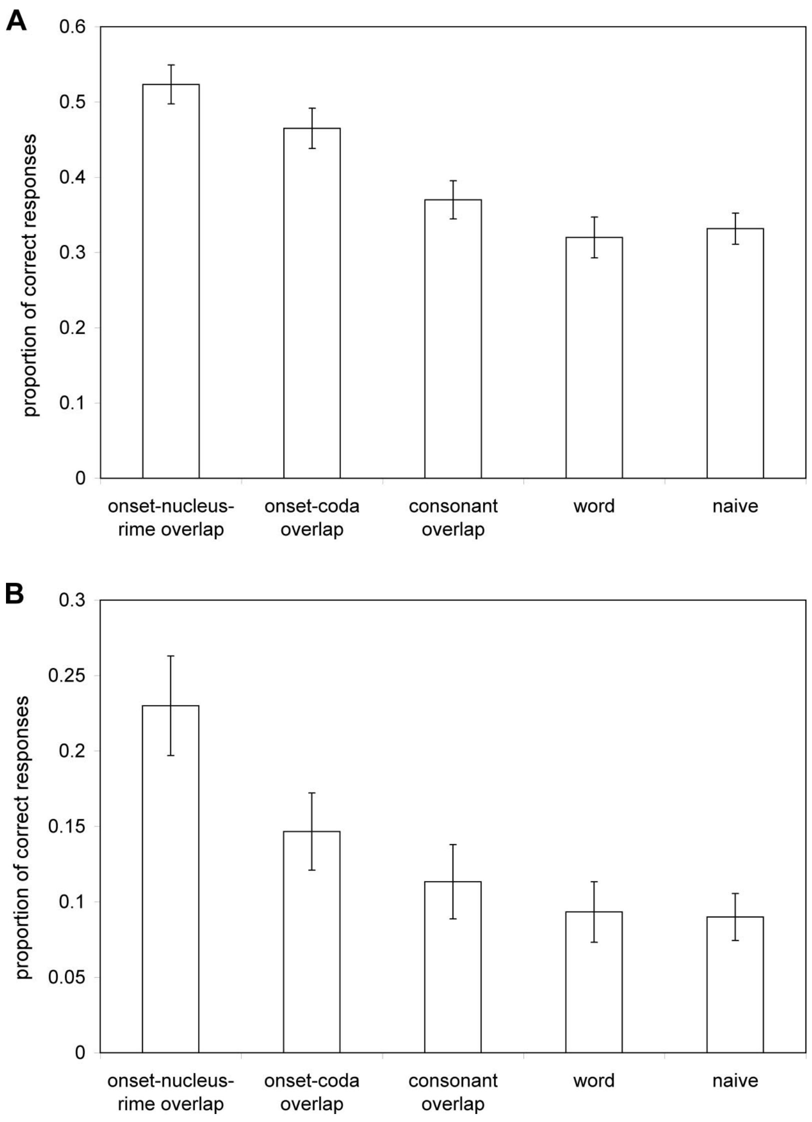

Figure 1. Experiment 1: Proportion of correct consonant identifications (A) and of correct word identifications (B) as a function of participants' training group. Error bars represent standard errors computed across participants.

Lexicality effect. Finally, we compared participants' performance when they had been trained with nonword stimuli vs. word stimuli when the similarity of the training and testing stimuli was equated between the two groups. Figure 1 indicates that word stimuli training resulted in slightly *worse* performance than nonword stimuli. As in the other analyses, we fitted a mixed logit model to our data. This model included training (word or nonword stimuli) and lexical frequency of the test word as fixed effects, as well as participants, items, and consonants (intercepts) as random effects. In this model, the nature of the training stimuli (word vs. nonword) made no significant contribution.

Word identification. Performance on word identification was analyzed to assess how the probability of generalizing the knowledge acquired at training to new words varied as a function of the training listeners received. The data are presented in Figure 1B. Apart from a lower level of performance, word identification across training groups patterned similarly with consonant identification: Greater phonetic overlap between training and testing stimuli resulted in better performance, and word identification following exposure to training words was marginally lower than that following exposure to training nonwords. To test for statistical significance, we modeled our data using similar models as those used to evaluate consonant identification.

Syllable-position match effect. First, we examined whether word recognition varied whether the consonants in the testing word had been heard at training in the same position or in the opposite position (i.e., the contrast between both the onset-nucleus rime overlap and the onset-coda overlap groups and the consonant overlap group). By contrast to what was observed on consonant identification, the match in consonant position had no significant contribution in predicting whether or not the entire word was correctly recognized, and neither did the lexical frequency of the testing word.

Vocalic context effect. To evaluate the contribution of the vocalic context in which the tested consonants were heard to word recognition, we compared the performance of the onset-nucleus rime overlap group with that of the onset-coda overlap group. For both groups, consonants at testing occurred in the same position as at training. However, these consonants were systematically flanked by the same vowel in the onset-nucleus rime overlap group only. To test whether the vowel match had a systematic contribution to consonant identification, we fitted a mixed logit model that included, as fixed effects, the vowel match and the frequency of the testing word, to the full data set and the subset that excluded the four testing words that shared their onset and nucleus with some of the training items in the onset-coda overlap group. In both of these models, with participants and items as random effects, the vowel context had a significant contribution to predicting word recognition, with a higher probability of correctly identifying a testing word when it overlapped with training stimuli by its onset-nucleus or rime than when it overlapped only by its onset or coda consonant (coefficient = 1.19, p = .006). The frequency of the testing word did not significantly predict word identification. However, there was a significant interaction between the vowel match and frequency (with a coefficient of −0.34, p = .050), such that the benefit associated with a vowel overlap decreased as the frequency of the testing word increased. This suggests that past experience with noise-vocoded stimuli becomes negligible to correctly identifying highly frequent testing words.

Effect of prior exposure with noise-vocoded speech. We examined the effectiveness of being trained to interpret noise-vocoded speech by listening to nonsense syllables (collapsing the three overlap groups), compared to receiving no training with this form of speech (i.e., the naïve group), by fitting the data to a logit mixed model with participants and items as random effects, and training status and test word frequency as fixed effects. The analysis revealed a significant contribution of training, with a coefficient of 1.54 (p = .0002). The odds of correctly identifying the testing word were 4.7 times greater when participants had received training with noise-vocoded speech than when they had not, even when this training was limited to hearing distorted nonsense syllables. There was also a significant interaction between training and the frequency of the testing words with a coefficient of −0.56 (p = .0006): As the frequency of the testing word increased, the advantage of having received noise-vocoded training (as opposed to no training) decreased. This result highlights the contribution of biases (or priors) to the word recognition in a highly distorted auditory signal, and how the contribution of these biases and that of the knowledge of acoustic-phonetic mapping acquired during training trade with one another.

Lexicality effect. We compared participants' performance when they had been trained with nonword stimuli vs. word stimuli when the similarity of the training and testing stimuli was equated between the two groups (thus, limited to the consonant-overlap groups). The simplest and best fitting logit mixed model included participants and items as random effects. In this model, the lexicality of the training stimuli made a significant contribution to predicting word identification at test (coefficient −1.35, p = .03): The odds of correctly identifying testing words were 3.9 times *smaller* for people trained on word than nonword stimuli. A significant interaction between lexicality and the testing stimuli's log frequency was also found (coefficient = .83, p = .0004), which indicated that, as the frequency of the testing word increased, the difference in performance between the listeners trained on word and nonword stimuli decreased.

Discussion

Experiment 1 examined the influence of the phonetic similarity between training and testing noise-vocoded stimuli on listeners' ability to generalize the knowledge acquired during training. We varied whether the testing words contained consonants that had appeared at training in the same syllable position or in the opposite, and whether these consonants were flanked by the same vowel or not. We found that consonants were more accurately identified when listeners had been exposed to them in the same syllable position during training than when they had been exposed to them but in a different position. This suggests that the adjustment to the noise-vocoded distortion is quite specific to the allophonic realization of these consonants and does not readily transfer to other allophones within the same phonemic category. The probability of correctly identifying the test word, however, was not significantly affected by the similarity, in syllable position, between its consonants and those heard during training. This result suggests that word identification may not be as sensitive a measure of listeners' generalization of their experience with noise-vocoded speech as consonant identification is. Indeed, word responses must reflect listeners' evaluation of the speech signal in terms of speech sounds and their knowledge of words and their priors. Evidence for this interaction was found in the apparent trade-off between the information that listeners have acquired during training and the priors they applied to the words they tried to recognize in the acoustic signal. The more frequent the testing word, the less of an impact prior knowledge on noise-vocoded speech had on people's ability to accurately identify the testing words. Although this trade-off was not unexpected, it is interesting to observe it in the context of relatively poor performances: Across all training groups, the testing word that was accurately identified most often was recognized only 44% of the time.

There was some indication that the vowel context in which the trained consonants had been encountered had some impact on subsequent consonant identification, but this effect was more modest. A possible explanation for the modest contribution of vowel context may hinge on its relatively crude manipulation in the present experiment. To assess the contribution of vowel context, we contrasted consonant-identification performance when the test consonant had appeared with the same vowel at training and performance when the test consonant had appeared with a different vowel, without further distinction. Some of these different vowels may have been acoustically quite similar to the training vowels, which in turn may have reduced the contrast between the two groups. We revisit the role of the vowel context in Experiment 4, where the match or mismatch between vowel at training and testing was more tightly controlled.

Before discussing Experiment 1's findings further, it is important to consider an alternative interpretation of the results. The influence of the syllable-position match on participants' performance may not reveal listeners' learning to interpret noise-vocoded position-specific acoustic patterns, but a tendency for participants to guess test words that adhered to the consonant distribution of the training items they received. According to this interpretation, the clear version of the training utterances led to response biases. To evaluate this alternative account of the present results, we focused on participants' incorrect consonant responses and recoded them as corresponding to one of the consonants people heard in the same position during training or not. The proportion of incorrect consonant responses that matched the training stimuli was 37% on average (varying across groups, between 35% and 42%). Thus, more often than not, participants (incorrectly) responded with words whose consonants did not adhere to the consonant distributions they had experienced during training. Furthermore, the small effect of vocalic context is not consistent with an account in which participants' responses were merely constrained by the consonant distribution heard at training. Thus, although we cannot rule out the presence of a bias that would incorporate the entire distribution of sounds from the training material, the data indicates that the influence of any possible bias must have been combined with listeners' analysis of the speech signal.

Experiment 1 also revealed that participants who had been exposed to nonsense distorted stimuli during training and their undistorted counterparts achieved better performance than participants who had not been exposed to the distorted form of speech (i.e., the naïve group), indicating that nonsense stimuli can provide substantial information on the nature of the noise-vocoded distortion. Furthermore, there was no evidence that being trained on word stimuli, as opposed to nonword stimuli, enhanced learning. The fact that listeners' ability to adjust to noise-vocoded speech was not facilitated by having been trained on words as opposed to nonwords is a direct challenge to the Davis et al. (2005) results and is in line with the more recent findings by this research group (Hervais-Adelman et al., 2008). However, we found a trend toward worse performance when listeners were trained on words than when trained on nonwords. This result raises the possibility that word and nonword training stimuli varied on other dimensions, in addition to their lexicality status. The acoustic similarity between the testing words and the training stimuli may have been greater for the nonword stimuli than for the word stimuli. In Hervais-Adelman et al., the two sets of stimuli were characterized by (roughly) the same number of occurrences of specific phonemes, with no attempt to match their position or vowel contexts. In Experiment 1, the two sets of stimuli were matched for the position of the consonants but not for the vowels that flanked the consonants. If the nonword training stimuli were acoustically more similar to the testing items than the word training stimuli were, this difference may have concealed the purported benefits from being exposed to words in training.

To address this concern, we replicated Experiment 1 with sets of word and nonword training stimuli that were tightly controlled for the nature and frequency of occurrence of their subcomponents. As in Experiment 1, training stimuli were CVC sequences that consisted of existing English words or nonsense syllables. Importantly, these word and nonword sequences were matched on segment and bigram frequencies. This new set of stimuli allowed us to assess the effect of the lexicality status of the training stimuli on learning under tightly controlled conditions.

Considering the influence of lexicality on learning to adjust to noise-vocoded speech is especially interesting in conjunction with assessing people's ability to generalize their experience with this form of speech. It is possible that being trained with distorted existing words enhances generalization because listeners may recognize the phonological categories to which the subparts of the words they hear belong; this categorization may not take place with one or two exposures to a novel string. Thus, Experiment 2 assessed how the nature of the training stimuli interacts with listeners' ability to correctly recognize testing consonants with various levels of similarity with training stimuli. We assessed the identification of consonants that had been present in training stimuli in the same or different syllable position and the identification of consonants that had not been heard at training.

Experiment 2

Method

Participants. Forty students from the University of Pennsylvania took part in this experiment for course credit. All were native speakers of English, and none had taken part in Experiment 1. They provided informed consent prior to the beginning of the experiment.

Materials. Following a procedure described by Sternberg, Wright, Knoll, and Monsell (1980), we constructed sets of eight CVC stimuli by selecting two onset consonants, two vowels, and two coda consonants which, when exhaustively combined, formed four existing English words and four nonsense syllables, to be used as training stimuli. Table 2 presents an example of such a set and its resulting words and nonwords.

The procedure used in constructing the stimuli was as follows. A list of the 2,545 monosyllabic, 3-phoneme long, words of English was generated, based on the Internet-accessible English Lexicon Project database (Balota et al., 2007). This list was hand-edited to exclude words that began or ended with a vowel, particularly infrequent words, and sequences that did not correspond to a single morpheme (e.g., *he's*). Six-segment sets (two onset consonants, two vowels, and two coda consonants) were then compiled provided that, when combined, they formed four words present in the edited list of English monosyllabic words, and four nonword sequences, that is, sequences that were not present in the CVC-word list and subsequently confirmed as nonexisting English words.

We selected 15 such six-segment sets. Together, the sets involved seven different onset consonants, 11 different vowels, and four different coda consonants. Each six-segment set generated 4 word and 4 nonword stimuli, for a total of 60 word stimuli and 60 nonword stimuli. When the same word or nonword stimulus was generated by combining the segments from more than one set, extra occurrences of the same stimulus were removed only if the balance in bigram frequencies between word and nonword stimuli was maintained. This procedure yielded 50 word tokens (comprising 39 distinct words) and 50 nonword tokens (comprising 38 distinct nonwords). Because of the stringent constraints on training-stimuli construction, the frequency with which each consonant was heard during training was allowed to vary. Note, however, that the frequency of consonant occurrence was perfectly

Table 2

Experiment 2: Examples of Testing Words as a Function of the Position of the Critical Consonant (i.e., in the Same Syllable Position as at Training or in a Different Syllable Position From That at Training) for Each Testing List

	Testing list A		Testing list B	
	Word training stimuli *gem joke deck dome*	Nonword training stimuli *dem doke jeck jome*	Word training stimuli *gem joke deck dome*	Nonword training stimuli *dem doke jeck jome*
Onset consonant: /d/, /dʒ/ vowels: /ɛ, o/ coda consonant: /k, m/	Same position	Different position	Same position	Different position
/d/	*dawn* *dull*			*loud* *hide*
/dʒ/		*lodge* *wage*	*join* *jeep*	
/k/	*hook* *pick*			*cape* *cop*
/m/		*mill* *mall*	*ham* *lamb*	

matched between the word and nonword training sets. Appendix B lists the complete set of training stimuli.

To construct the testing stimuli, we selected, from the set of consonants used in the training stimuli, consonants that are permissible in both onset and coda positions in English. These "critical" consonants were a subset of the trained consonants that occurred at onset (namely, /b/, /d/, /f/, and /dʒ/) and all of the trained consonants that occurred at coda (/k/, /m/, /s/, and /t/). Based on these consonants, we constructed two distinct lists of CVC testing words. In these lists, we varied which of the onset and coda consonants appeared in the same syllable position as, or in a different syllable position from, the position they occupied in the training stimuli. The syllable-position factor was varied within consonants (as opposed to between consonants, as in Experiment 1) because of the small number of consonants involved here. List A consisted of 16 CVC testing words, each one containing one of the critical trained consonants. For half of the 16 testing words, the critical consonant was at onset, and for the other half, at coda. For half of the 8 words with the training consonant at onset, this consonant was in the same position as in training, and the other half, in the other position (see Table 2). For example, in List A, the consonant /d/, which was one of the onset consonants in the training stimuli, appeared at onset in two testing words (e.g., *dawn, dull*) and never at coda. The other consonant of each testing word (e.g., *n* in *dawn* and *l* in *dull*) was one of the training consonants other than the critical consonants or a new consonant. Thus, each testing word contained only one critical consonant, and only the recognition of this consonant was coded. This circumvented the lack of independence between the recognition of each of the two consonants within the same word.

List B was created to mirror the distribution of the training consonants over the testing words. It consisted of 16 different testing words that contained the same critical consonants as in List A, but in the opposite syllable position. For example, the consonant /d/ appeared at coda in two testing words of List B (i.e., *loud, hide*) and never at onset. These two lists of 16 testing words with trained consonants were matched on mean word frequency. Except for one item (in List A), testing words never shared the same onset-nucleus or rime with training words. Thus, across the two testing lists, we assessed listeners' ability to recognize a trained consonant when it appeared in the same syllable position as in training (but flanked by a different vowel) or in a different position. The recognition of each consonant was assessed twice, in two different testing words.

Finally, a set of 7 consonants (/g/, /r/, /θ/, /tʒ/, /ʒ/, /v/, and /z/) that never appeared in the training stimuli (and were not used as filler consonants in the other testing words) was used to construct 28 additional CVC testing words. Each of these "untrained" consonants appeared in four different words, twice at onset and twice at coda. Here again, the other consonant of each of these 28 testing words consisted of one of the filler consonants at training or a new consonant. Lists A and B and the 28 testing words with untrained consonants are presented in Appendix B.

All training and testing stimuli were read by the second author and recorded directly onto a computer (sampling at 22 kHz with 16-bit resolution). Each stimulus was then edited and a noise-vocoded version of each utterance was generated, using the same procedure as in Experiment 1.

Procedure and design. Participants were randomly assigned to one of four groups. In two of these groups, participants were first exposed to the 50-word training stimuli, each one presented twice, yielding 100 training tokens. In the other two groups, participants were exposed to the 50-nonword training stimuli, each one presented twice, yielding 100 training tokens. Instructions to participants explicitly stated that these stimuli were, depending on the group, real, existing English words or made-up words. The order with which the training stimuli were presented was varied by creating four subsets of 25 stimuli and randomizing the order of presentation of each subset across participants. As in Experiment 1, each training token was presented three times successively, first in its distorted version, then in its original version, and finally in its distorted version again. Following the training phase, participants completed the testing phase. Each of the testing words was presented only once, in its distorted version, and participants were instructed to guess which word they heard and to enter it onto the keyboard. They moved to the next testing word by pressing the space bar.

Half of the word-trained participants were tested on the 16 testing words (containing familiar consonants) from List A, the other half, on the 16 testing words from List B. Likewise, half of the nonword-trained participants were tested on the 16 testing words from List A, the other half, on the 16 testing words from List B. All participants were tested on the 28 testing words with unfamiliar consonants. The order of presentation of the 44 testing words was randomized.

Results

Consonant-identification accuracy was evaluated only for the critical consonant of each testing word, using the same stringent coding scheme as in Experiment 1.

The role of the training regime was evaluated in two separate analyses. The first analysis focused on the accuracy on trained consonants only, as a function of their syllable-position match between training and testing. The second analysis compared accuracy to trained and untrained consonants. Figure 2 presents the proportion of correct consonant identifications as a function of the consonants' syllable-position match between training and testing, and whether participants were trained on word or nonword stimuli (Figure 2A). As apparent on the figure, trained consonants that appeared in the same position as at training were correctly identified more frequently than trained consonants that appeared in the opposite position. There was no influence of the nature of the training stimuli (word vs. nonword).

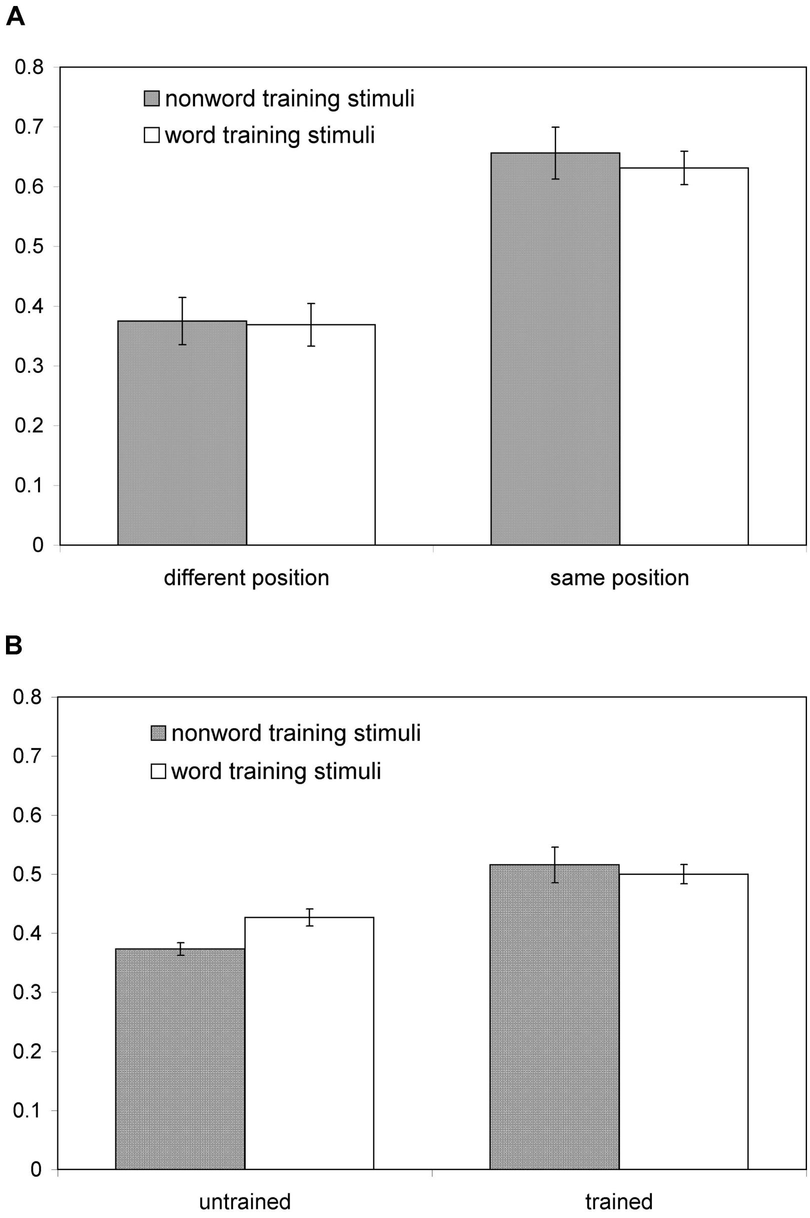

Figure 2. Experiment 2: Proportion of correct consonant identifications as a function of (A) the match in syllable position between the trained consonants at training and testing and participants' training stimuli (word vs. nonword) and (B) the familiarity of the testing consonants (trained vs. untrained) and participants' training stimuli (word vs. nonword).

To confirm this picture, we fitted a mixed logit model to the data that specified, as fixed effects, training and syllable-position match (their main effects and interaction). The simplest model, i.e., the model with the fewest random terms with the best fit to the data, was the model that estimated different intercepts for each item (i.e., testing word) as random effect. (Adding estimates for intercepts for each participant or consonant as random effects did not improve the fit significantly.) In this simplest model, the syllable-position match between the consonants at training and testing had a significant impact on the probability of correctly identifying the consonants, with a coefficient of 1.80 ($p = .02$). This coefficient indicates that the odds of correctly identifying a testing consonant presented in the same position as in training were 6 times higher than the odds of correctly identifying a testing consonant that had been heard in a different position at training. The training regime made no significant contribution to predicting consonant identification. The interaction between training regime and syllable-position match was also not significant. Thus, there was no evidence that recognition of trained consonants was affected by having been trained to process noise-vocoded speech corresponding to existing words or nonsense syllables.

The second analysis compared the identification of trained and untrained consonants as a function of the training stimuli (words vs. nonwords). Figure 2B presents the proportion of correct consonant identifications as a function of the familiarity of the consonants (trained vs. untrained) and the training stimuli participants received. Overall, trained consonants (collapsing across consonants appearing in the same position as at training and in a different position) were correctly identified more frequently than untrained consonants were. The nature of the training stimuli had no impact on the recognition of trained consonants (as the first analysis showed), but did modulate the recognition of untrained consonants: Participants were more likely to correctly identify untrained consonants if they had been trained with word stimuli than with nonword stimuli. As before, we tested this by fitting a mixed logit model to the data. In terms of random effects, the simplest model included estimates of intercepts for each item and for each consonant. Fixed effects were training regime (nonwords vs. words) and consonant familiarity (trained vs. untrained consonant) and the interaction between the two. In this model, there was no main effect of familiarity with the consonant at testing (probably owing to the fact that the position match of the trained consonant, a factor that this analysis could not include, contributed substantially to its recognition), a main effect of training with a coefficient of 0.44 ($p = .01$), and more notably, a significant interaction between with a coefficient of -0.56 ($p = .04$). The interaction coefficient indicates that the odds of correctly identifying a consonant were 1.6 times greater if participants had been trained with real words. However, the main effect of training is not very meaningful on its own, given the significant interaction with familiarity. The effect of training was modulated by which type of consonants people identified: While the effect of being trained on words vs. nonwords *increased* the odds of correctly recognizing untrained consonants, it *decreased* the odds of correctly recognizing trained consonants, although only modestly so, as observed in the results from the first analysis. Thus, there is some evidence that having being trained to decode noise-vocoded speech with word stimuli enhanced listeners' ability to generalize their knowledge to different, untrained consonants, compared to receiving training with nonsense syllables.

Discussion

Experiment 2's results confirmed the importance of the similarity that training and testing consonants share, expressed in terms of syllable position. Generalizing experience with consonants at training was superior when the consonants were presented in the same position than in a different position. Experiment 2's results also showed an intriguing effect of the training regime. When provided with the clear, undistorted versions of the training stimuli, people learned to recognize trained consonants equally well whether the training stimuli consisted of existing words or nonsense syllables. However, having been trained on existing words appears to have provided listeners with the greater ability to generalize to different, untrained consonants. We return to the implication of this finding in the general discussion.

Experiments 3 and 4 examined how experience with consonants at training later generalizes to a different speaker's productions. As pointed out earlier, the ability to generalize across speakers is a necessary condition for a robust recognition system, but one that has proved difficult to achieve in automatic-speech recognizers (see Huang, Acero, & Hon, 2001). Despite the remarkable ability of human listeners to understand any speaker of their language, there is good evidence that past exposure with a talker facilitates the ease and speed with which listeners process that talker's utterances (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). This finding suggests that the acoustic aspects that differ across talkers play a role in speech recognition. Studies that have focused on listeners' adaptation to artificially distorted speech, however, have reported evidence that listeners can generalize their learning to the speech of a novel talker that has undergone similar distortion (Dupoux & Green, 1997). Results from Experiments 1 and 2 have revealed limitation in listeners' generalization: People's ability to recognize trained consonants was greatly influenced by the phonetic similarity between their past experience with that consonant and the realization of that consonant at test. Here, we ask if speaker generalization is characterized by the same limitation or if it operates more readily. Although much spectral information is lost in noise-vocoded speech, there is some evidence that people can reliably discriminate a male talker from a female talker even when most of the spectral information in the signal has been removed (Fu, Chinchilla, Nogaki, & Glavin, 2005). Furthermore, speaker identity may be preserved in other acoustic cues that remain relatively intact in noise-vocoded speech. Here, we ask whether the presence of such cues may constrain listeners' subsequent ability to generalize their experience to different stimuli.

Experiment 3

Method

Participants. Forty-eight students from the University of Pennsylvania took part in this experiment for course credit. All were native speakers of English, and none had taken part in Experiment 1 or 2. They provided informed consent prior to the beginning of the experiment.

Materials. Two sets of training stimuli were constructed. All stimuli were existing English CVC words. The two sets differed in the position that their consonants held. In set A, the onset consonants were /d/, /k/, /m/, /p/, /r/, and /v/ and the coda consonants were /b/, /f/, /l/, /t/, /n/, and /g/. In set B, the same consonants were used but in opposite syllable positions. The two training sets were further matched on the vowel that flanked each consonant. For example, the word *cab* in set A, in which /k/ occurs at onset and /b/, at coda, both flanking the vowel /æ/, was matched to the word *back* in set B, in which /k/ occurs at coda and /b/, at onset and both flanking /æ/. These constraints could be met either within a single word pair (as in *cab*/*back*) or within word quadruplets (e.g., with *calf* and *Dan* in set A, and *fad* and *knack* in set B). Each consonant appeared in four different training words, yielding a total of 24 words in each set. Sets were matched in frequency as closely as possible, although set B's words tended to be more frequent. (Unlike set B, set A contained a proper name, *Dan*.) Appendix C lists each set's stimuli.

A single set of 96 testing words was constructed. For each of the 12 trained consonants, 8 CVC words that contained that consonant were selected, with the consonant occurring at onset for half of the words and at coda, for the other half. The other consonant of each testing word was never one of the trained consonants, and the vowel that flanked the trained consonant had never co-occurred with that consonant at training. Each set of four words per trained consonant per syllable position was further divided in half, with one half assigned to one speaker and the other half, to the other speaker. Table 3 illustrates how the testing words related to the training words. The complete set of testing words is presented in Appendix C.

All training and testing stimuli were read by two female speakers, one of them being the second author (RM), and directly recorded onto a computer (sampling at 22 kHz with 16-bit resolution). Each word was then edited and a noise-vocoded version of each utterance was generated, using the same procedure as in Experiment 1.

Design and procedure. Participants were randomly assigned to one of four groups, varying which training set they heard and which speaker the recording originated from. The training part of the study proceeded as follows. Each participant was exposed to one of the two training sets of 24 words (set A or set B). To increase the amount of training people received, the entire set was repeated twice. For each presentation of the set, the order of the training words was randomized for each participant. As in the previous experiments, each training word was presented three times successively, first in its distorted version, then in its original version, and finally in its distorted version again. People were instructed that they would hear English words (and, for people exposed to set A, that some of them could be proper names). Their task was to listen carefully. Following the training phase, participants completed the testing phase. Instructions specified that some of the testing words had been produced by the same speaker as in training and others, by a different speaker. Each of the 96 testing words was presented once, in its distorted version only, and participants were instructed to identify the word they heard and to enter it onto the keyboard. They moved to the next testing word by pressing the space bar.

Results and Discussion

From participants' responses to each testing word, the accuracy of the consonant of interest was assessed, using the same criterion as in previous experiments. Figure 3 presents the proportions of correctly identified consonants as a function of their syllable-position status (in the same position at testing as the position they occupied at training or in a different position) and the speaker status (produced by the same speaker as the speaker participants were trained on or by a different speaker). As apparent in the figure, people accurately identified the test consonants more often when the consonants occurred in the same syllable position as the one they occupied at training and when they had been pronounced by the same speaker as the one people had been trained on. We note that, when expressed in proportions, the data seem to reveal an interaction between the two factors, with a greater effect of the speaker match on the identification of consonants that appeared in a different position than the one they occupied at training than on the identification of consonants that appeared the same position. This impression is mistaken, however. Indeed, once the propor-

Table 3
Experiment 3: Illustration of the Design With Examples of the Training and Testing Stimuli

	Training group							
Variable	Set A speaker *SD*		Set A speaker RM		Set B speaker *SD*		Set B speaker RM	
Example of training word	*cab*$_{SD}$		*cab*$_{\mathrm{RM}}$		*back*$_{SD}$		*back*$_{\mathrm{RM}}$	
Example of testing words	Position status	Speaker status	Position status	Speaker status	Position status	Speaker status	Position status	Speaker status
cage$_{SD}$	Same	Same	Same	Different	Different	Same	Different	Different
coach$_{\mathrm{RM}}$	Same	Different	Same	Same	Different	Different	Different	Same
hike$_{SD}$	Different	Same	Different	Different	Same	Same	Same	Different
joke$_{\mathrm{RM}}$	Different	Different	Different	Same	Same	Different	Same	Same
booth$_{SD}$	Same	Same	Same	Different	Different	Same	Different	Different
beach$_{\mathrm{RM}}$	Same	Different	Same	Same	Different	Different	Different	Same
web$_{SD}$	Different	Same	Different	Different	Same	Same	Same	Different
hub$_{\mathrm{RM}}$	Different	Different	Different	Same	Same	Different	Same	Same

Note. The position status of the underlined consonant in each of the testing words captures whether the consonant had appeared in the same syllable position at training ("same") or in the opposite position ("different"). The speaker status of the underlined consonant in each testing word captures whether the consonant had been heard at training pronounced by the same speaker ("same") or by a different one ("different"). Subscripts "RM" and "SD" refer to the initials of the speaker whose production was used at training or testing.

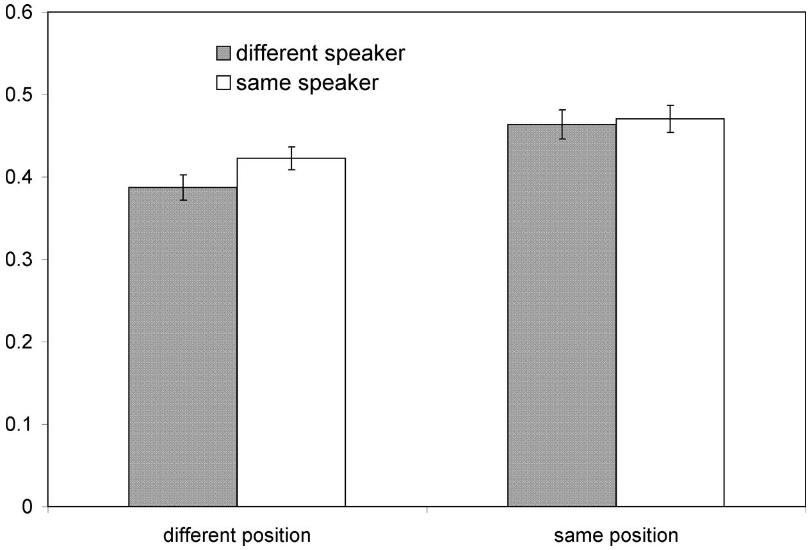

Figure 3. Experiment 3: Proportion of correct consonant identifications at test as a function of their syllable-position status (i.e., in the same position as in training or in a different position) and their speaker status (pronounced by the same speaker as in training or by a different speaker).

tions are expressed into odds, the change in consonant identification when the consonants originated from the same speaker, as opposed to a different speaker, results in a 1.1-fold increase whether the consonant was in the same position as at training (from 38.7% to 42.3%) or in a different position (from 46.3% to 47.0%) (see Jaeger, 2008, for further discussion).

To evaluate the significance of the observed differences and the interaction between these two factors, the data were analyzed by fitting mixed logit models to the data. The fixed effects were the main effects of the syllable-position status of the consonant (same as at training or different) and of the testing word's speaker status (same as at training or different), as well as the interaction between these two factors. The model that fitted the data best included three random effects, accommodating specific intercepts for individual participant, testing word, and consonant. In this model, the syllable-position status of the consonant significantly influenced the probability of correct identification, with a coefficient of 0.54 ($p < .00001$): The odds of correctly recognizing a consonant were 1.7 times greater when the consonant occurred in the same syllable position as in training than when it occurred in the opposite position. The speaker status also had a significant contribution to predicting consonant-recognition performance, with a coefficient of 0.25 ($p = .02$): The odds of recognizing a consonant were 1.3 times greater when the consonant had been heard pronounced by the same speaker at training than when it had been heard pronounced by a different speaker. Finally, there was no significant interaction between these two factors, suggesting additivity in each factor's contribution. The probability of demonstrating ability to generalize past experience with a consonant decreases with the number of factors to generalize over. When a consonant at test shared both its syllable position with training stimuli and the speaker that produced them, the least generalization was required and performance was best. When a consonant at test differed from tokens heard at training in terms of both position and speaker, performance was the poorest. When a consonant differed on only one of those dimensions, performance was intermediate.

Experiment 3's results confirmed the importance of the similarity between training and testing consonant tokens in predicting consonants' recognition. In addition to replicating the effect of syllable position, Experiment 3 suggests that the acoustic characteristics that distinguish the realizations of the same word from two different speakers influence listeners' ability to correctly identify consonants. This speaks to listeners' limited ability to generalize their learning to decode noise-vocoded speech to acoustically different stimuli. Our finding that speaker identity plays a significant contribution to generalization is all the more remarkable because the noise-vocoded transformation removes a lot of the spectral information that differ between talkers. The voices of our speakers, both female, were characterized by quite similar pitch: The mean F0 value, computed over the training stimuli were 214 Hz for speaker RM and 236 Hz for speaker *SD* Presumably, the articulatory differences between speakers that affected the temporal characteristics of their speech must have played a major role.

Experiment 4 aimed to confirm the influence of speaker's identity on listeners' ability to generalize their experience with noise-vocoded speech to new stimuli by contrasting two speakers that differed in their pitch level more than the two speakers we contrasted in Experiment 3. We examined whether the contribution of the speaker identity to consonant identification is greater if the speakers' voices are more distinguishable once their speech has been noise-vocoded. We selected two speakers whose pitch would fall within two different frequency bands, as defined by the noise-vocoded transformation and over which the amplitude envelope is extracted to modulate band-limited noise. Experiment 4 also investigated listeners' ability to generalize their experience with a speaker's vowels to another speaker's vowels. Examining vowel perception is particularly relevant to cross-talker generalization because, to a large part, speakers' identity is encoded in the spectral composition of segments like vowels. Finally, we revisited the question of whether consonant recognition is sensitive to the similarity in the vocalic context in which it was heard at training.

Experiment 4

Experiment 4 examined listeners' generalization of their experience with noise-vocoded speech to the speech of a different talker. Because the talkers' identity is especially salient in the acoustic characteristics of vowels, the recognition of vowels, in addition to that of consonants, was assessed here. We also examined the influence of vowel context on the recognition of consonants: The tested consonant had always been heard at training in the same position. What we varied was whether the consonants had been heard with the same flanking vowels or flanked with vowels that were different *and* acoustically distant from the present contextual vowel. This manipulation was achieved by exposing participants to training stimuli that contained either front or back vowels. The designation "back" and "front" refers to the position of the tongue in the vocal tract during the articulation of the vowels, and this articulatory configuration has acoustic consequences: In acoustic space, by and large, formant frequencies of front vowels are closer to one another than they are to those of back vowels, and formant frequencies of back vowels are closer to one another than they are to those of front vowels. By controlling the type of vowels people heard at training, we ensured that the vowels at testing were either the same as in training or as different as possible.

Method

Participants. Forty-four students from the University of Pennsylvania took part in this experiment for course credit. All were native speakers of English, and none had taken part in any of the previous experiments. They provided informed consent prior to the beginning of the experiment.

Stimuli. Two sets of training stimuli were constructed. The sets consisted of existing English CVC words that began with the consonants /r/, /b/, /p/, or /n/ and ended with the consonants /l/, /d/, /k/, or /m/. The sets differed only in the words' vowels, with set A containing back vowels, and set B, front vowels. For example, set A included the words *roam* and *rum*, both containing back vowels, and set B included the words *rim* and *ram*, both with front vowels. In each set, four different vowels were used and each one occurred once with each onset and coda consonant, yielding a total of 16 training words per set. Appendix D lists the words for each set.

For each of the 32 training word (e.g., *roam*), we selected as testing words two existing English CVC words, one that shared its onset consonant and vowel with the training word (i.e., *rose*) and one that shared the training word's vowel and coda consonant (i.e., *home*) (see Table 4, for an illustration). The other consonant of each testing word was never one of the consonants that appeared during training. This yielded 64 testing words. Half of them contained an onset or coda consonant that had appeared with the same vowel as in set A (i.e., a back vowel), and the other half, an onset or coda consonant that had appeared with the same vowel as in set B (i.e., a front vowel). Each of the 8 consonants was tested in 8 different testing words, 4 with back vowels and 4 with front vowels. Note that, contrary to the preceding experiments, the tested consonants had always appeared at training, and in the same syllable position. What varied here was whether the (onset or coda) consonant had appeared with the same vowel or not.

The 32 training words and 64 testing words were read by two speakers, one female (RM, the second author) and one male (CN) and recorded directly onto a computer. Each word was then edited. The mean F0 value on the 32 training words was 123 Hz for speaker CN and 233 Hz for speaker RM. Because the first frequency band involved in the noise-vocoded transformation comprises frequencies ranging from 50 and 229 Hz, speaker CN's pitch was likely to fall within the first band while speaker RM's pitch was more likely to fall within the second band. This maximized the chance that differences in pitch between the two speakers would be maintained in the distorted utterances. The recordings of 32 testing words produced by speaker RM and of the remaining 32 produced by speaker CN were selected. These two sets of testing words were balanced for the position of the tested consonants (at onset or coda)

Table 4
Experiment 4: Illustration of the Design With Examples of the Training and Testing Stimuli

	Training group							
Variable	Set A speaker RM		Set A speaker CN		Set B speaker RM		Set B speaker CN	
Example of training word	$roam_{\mathrm{RM}}rum_{\mathrm{RM}}$		$roam_{\mathrm{CN}}rum_{\mathrm{CN}}$		$rim_{\mathrm{RM}}ram_{\mathrm{RM}}$		$rim_{\mathrm{CN}}ram_{\mathrm{CN}}$	
Example of testing words	Vowel status	Speaker status	Vowel status	Speaker status	Vowel status	Speaker status	Vowel status	Speaker status
$\underline{r}ush_{RM}$	Same	Same	Same	Different	Different	Same	Different	Different
$\underline{r}ose_{\mathrm{CN}}$	Same	Different	Same	Same	Different	Different	Different	Same
$gu\underline{m}_{\mathrm{RM}}$	Same	Same	Same	Different	Different	Same	Different	Different
$ho\underline{m}e_{\mathrm{CN}}$	Same	Different	Same	Same	Different	Different	Different	Same
$\underline{r}idge_{\mathrm{RM}}$	Different	Same	Different	Different	Same	Same	Same	Different
$\underline{r}ag_{\mathrm{CN}}$	Different	Different	Different	Same	Same	Different	Same	Same
$ha\underline{m}_{\mathrm{RM}}$	Different	Same	Different	Different	Same	Same	Same	Different
$whi\underline{m}_{\mathrm{CN}}$	Different	Different	Different	Same	Same	Different	Same	Same

Note. The vowel status of the underlined consonant and of the vowel in each of the testing words captures whether the consonant had appeared flanked by the same vowel at Training ("same") or not ("different"). The speaker status of the testing word captures whether the consonant or the vowel had been heard at training Pronounced by the same speaker ("same") or by a different one ("different"). Subscripts "RM" and "CN" refer to the initials of the speaker whose production was used at training or testing.

and for the nature of the vowel (back or front). The frequency with which each consonant occurred within each set was also balanced.[3]

Design and procedure. Participants were randomly assigned to one of four groups, depending on which training set they heard and which speaker the recording originated from (see Table 4). By varying the set and the speaker participants were trained on, we varied, across participants, the status of each of the testing word's critical consonant. As illustrated in Table 4, the onset consonant *r* in the testing word *rush* produced by Speaker RM contained the same vowel as that present in one of the training words and spoken by the same speaker for participants assigned to the group Set A speaker RM. However, the status of that consonant was different for participants who had been assigned to the other training groups. For instance, for people trained on Set A Speaker CN, the consonant *r* in *rush* contains the same vowel as at training but spoken by a different speaker. Thus, each testing word's similarity with the training stimuli in terms of vowel context and speaker varied as a function of the training group participants were assigned to.

The training part of the study proceeded as follows. Each participant was exposed to one of the two training sets of 16 words (set A or set B). To increase the amount of training people received, the entire set was repeated four times. The order of the 16 trials in each block was randomized for each participant. As in the previous experiments, each training word was presented three times successively, first in its distorted version, then in its original version, and finally in its distorted version again. People were instructed that they would hear English words, some of them being proper names. Their task was to listen carefully. Following the training phase, participants completed the testing phase. Instructions specified that some of the testing words had been produced by the same speaker as in training and others, by a different speaker. Each of the 64 testing words was presented once, in its distorted version only, and participants were instructed to identify the word they heard and to enter it onto the keyboard. They moved to the next testing word by pressing the space bar. The order of the testing trials was random and the same for all participants.

Results and Discussion

From participants' responses to each testing word, the accuracy of the consonant of interest was assessed, using the same criterion as in previous experiments. The accuracy of the vowel in participants' responses was also coded. In the rare cases where participants' answer contained more than one vowel (e.g., reporting *behave* upon hearing the testing word *pave*), the vowel was coded as correct if the vowel in the stressed syllable was the same as the testing word's vowel.

Figure 4 presents the proportion of correct consonant identifications (Figure 4A) and correct vowel identifications (Figure 4B). As apparent on (A), a given consonant was correctly identified more often when it had been heard at training in the same vowel context than in a different context. Whether the testing word had been pronounced by the same speaker as that trained on or not, however, had little impact on consonant recognition.

As in the previous experiments, these data were analyzed by fitting mixed logit models with speaker status, vowel status, and their interaction as fixed effects. The simplest and best fitting model included participants and consonants as random effects (intercepts). In this model, the vowel-context status significantly influenced the probability of correctly identifying the consonant, with a coefficient of 0.3 ($p = .03$): The odds of correctly recognizing a consonant were 1.3 times greater when the consonant had been heard in the same vowel context at training than in a different one. The speaker status, on the other hand, made no significant contribution to predicting consonant recognition, and there was no significant interaction between the two fixed effects.

The same picture emerged from the analysis of vowel identifications. The data are presented in Figure 4B. Vowels were correctly identified more frequently when they had been presented during training than when they were different from those trained on, with little influence of whether those vowels originated from the same speaker or a different speaker from the training speaker. The simplest and best fitting model, which included participants and items as random effects, revealed a significant contribution of vowel status, with a coefficient of .32 ($p = .01$): The odds of correctly identifying vowels at testing were 1.4 times greater when the vowel had been heard at training than when it had not. The speaker status, on the other hand, made no significant contribution to predicting vowel recognition, and there was no interaction between the two fixed effects.

Thus, Experiment 4's data revealed no significant benefit associated with processing speech that originated from the same speaker as at training: The probability of correctly identifying consonants and vowels in the testing words was the same whether these testing words had been produced by the same speaker as the one people received training on noise-vocoded speech or by a different speaker. This result is in contrast with Experiment 3's finding, despite having selected two speakers who contrasted more in some of their voice characteristics, namely their pitch and other features associated with speaker's gender.

To help interpret Experiment 4's results in the light of Experiment 3's, we conducted a control study that assessed listeners' ability to recognize the speaker they heard at training when hearing the test distorted utterances for both Experiments 3 and 4's stimuli. Twenty students from the same subject pool were recruited. Eight of them were tested on Experiment 3's stimuli, the other 12, on Experiment 4's stimuli. (Because the number of testing utterances differed between Experiments 3 and 4, the number of students tested on each stimulus set differed to equate the total number of observations collected in each group.) Participants were randomly assigned to one of the training lists. Their task was to listen to each of the training utterance carefully (presented in its distorted, clear, then distorted version) in anticipation of their assignment at testing, namely deciding whether the test utterance (presented once and in its distorted version) had originated from the same speaker as at training or from a different speaker. The procedure was identical to that used in Experiments 3 and 4, except that, after each test utterance, people pressed the key "S" on the computer keyboard when they judged the utterance to have been originally produced by the same speaker as at training, and the key "D",

---

[3] Because of an experimenter error, five of the eight testing words containing the onset consonant *n* were from speaker RM (and the remaining three, from speaker CN), and five of the eight testing words containing the onset consonant *b* were from speaker CN (and the remaining three, from speaker RM).

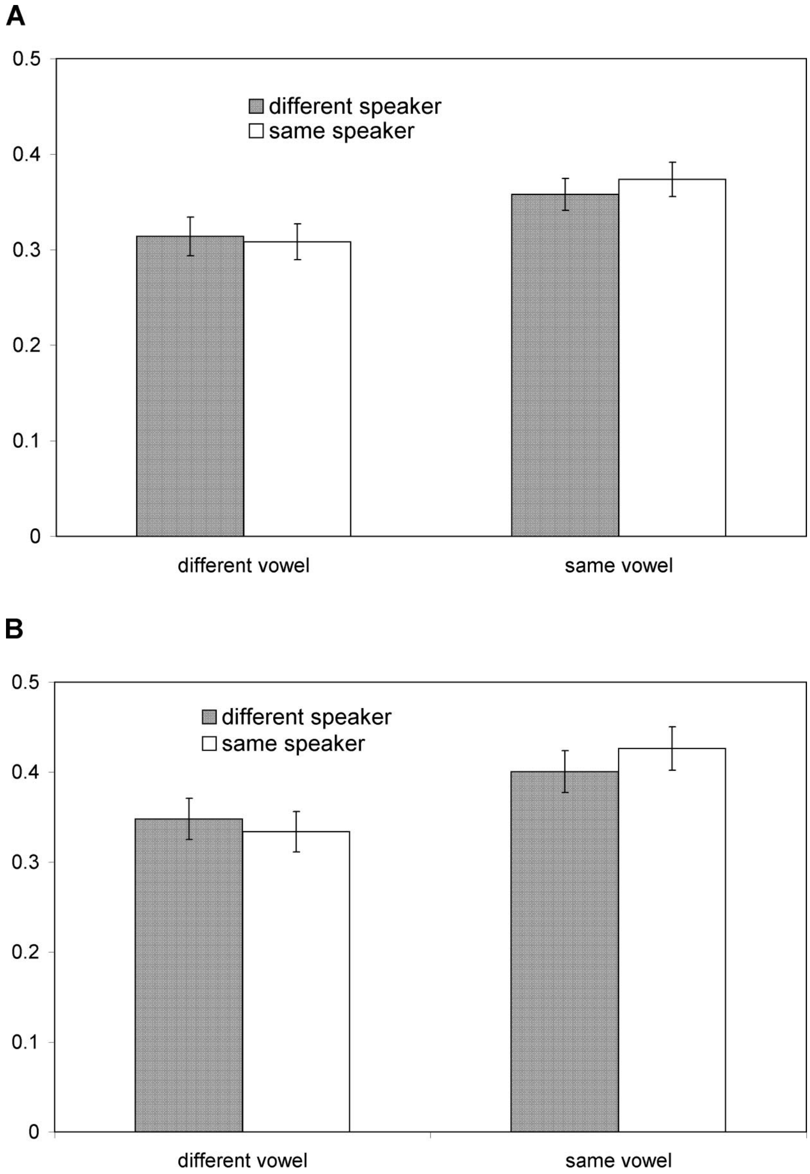

Figure 4. Experiment 4: Proportion of correct consonant identifications (A) and correct vowel identifications (B) at test as a function of whether or not the training words contained the same vowel and whether participants had been trained with stimuli pronounced by the same speaker or a different speaker.

when they judged the utterance to have been produced by a different speaker.

Participants who were exposed to Experiment 3's training stimuli and ask to recognize the speaker out of two female speakers responded correctly 52% of the time. Participants who were exposed to Experiment 4's training stimuli and ask to recognize the training speaker out of one female and one male speaker responded correctly 62% of time. A multilevel logistic regression analysis conducted on the recognition judgments made on Experiment 3's stimuli revealed that participants' performance did not differ from 50%. By contrast, the same analysis conducted on the judgments made on Experiment 4's stimuli revealed a significant intercept (with a coefficient of 0.56, $p < .001$), confirming that participants' performance was better than chance at recognizing the training speaker. Data from all participants were entered into a large multilevel logistic regression to compare performances on the two kinds of stimuli. Participants' ability to recognize the training speaker was significantly better when the two speakers involved at testing were of a different gender (i.e., Experiment 4's stimuli) than when they were of the same gender (i.e., Experiment 3's stimuli): People were 1.5 times more likely to correctly recognize the trained speaker when the two speakers to discriminate were of a different gender than when they were of the same gender (coefficient 0.42, $p < .05$).

This control study confirms our intuition that noise-vocoded utterances from speakers of different gender were easier to tell apart than utterances from speakers of the same gender. Importantly, the ability to consciously discriminate the speakers had little bearing on listeners' ability to generalize their learning to novel utterances. Indeed, greater cross-speaker generalization was observed on stimuli where speakers were easier to discriminate. We return to this point in the general discussion.

Finally, Experiment 4 showed that vowels that had been heard during training were more readily recognized than untrained vowels. Perhaps more interestingly, whether the testing word's vowel had been trained or not affected the recognition of its adjacent consonant. This result confirms with more tightly controlled stimuli the trend that we observed in Experiment 1. It suggests that people's ability to generalize beyond the training stimuli they received is constrained by the acoustic similarity between these stimuli and novel stimuli that a description in terms of phonetic categories (phonemes, allophones) cannot capture.

General Discussion

The present study examined how exposure to noise-vocoded utterances enables listeners to successfully generalize their experience to novel stimuli, and how phonetic and lexical factors may constrain this generalization. In each of the experiments reported here, participants were trained to interpret noise-vocoded CVC stimuli by hearing the original recording of each utterance along with its distorted counterpart. At testing, different stimuli were presented. The first experiment established that participants who received training were better able to accurately identify testing words and their subcomponents than people who had not had any past experience with noise-vocoded speech. Because testing stimuli were recordings of different words than those used at training, this demonstrates people's ability to generalize beyond the specific tokens they heard. However, this ability was greatly modulated by the degree to which testing stimuli phonetically resembled training stimuli. Testing stimuli that shared the same onset and nucleus or the same rime with the training stimuli were easier to recognize than stimuli that shared only their onsets or codas, and testing stimuli that shared consonants with the training stimuli were easier to recognize when these consonants had occurred in the same syllable position than in the opposite position. Thus, the present study highlights the fact that listeners' ability to successfully generalize their past experience increases as the similarity between the trained utterances and novel instances increases.

Our tests of cross-talker generalization provided an intriguing picture. In the two studies that manipulated whether the utterances had been produced by the same talker at testing as at training or by a different talker, the results pertaining to the contribution of this factor on generalization were inconsistent. In fact, the study that did not show an effect of the speaker match was the one where the two speakers were easier to discriminate, based on their distorted utterances. In that study, listeners generalized their experience with the consonants and vowels they heard at training to the same degree, irrespective of whether these consonants and vowels had been pronounced by the same speaker or a different speaker.

It is important to point out that the designs of the present studies enable us to comment on people's performance in relative terms only. Thus, for instance, Experiment 3's finding showed that having been trained to interpret one of the talker's distorted utterances leads to better performances on utterances from the same talker than from a different talker. These results do not demonstrate that people didn't generalize their training experience to utterances from a different talker, nor can they differentiate between a same-talker benefit and a different-talker penalty (whereby listeners' generalization of their learning experience had a detrimental effect on their performance). Comparison with the performance of a group of listeners who had received no prior exposure with the clear or distorted utterances of either of the talker heard at testing may help evaluate listeners' absolute performance in the same and different-talker conditions.

Before commenting on the theoretical implications of our findings, it is important to evaluate the possibility that the observed differences in the performance of participants who received different training may merely reflect participants' biases to offer responses at testing that resembled the clear stimuli heard at training. Such a bias may account for better consonant identification when the position of the consonant and its vocalic context matched that of the training stimuli than when they did not. However, a bias account fails to explain the speaker-match effect observed in Experiment 3. Participants were more accurate at identifying the consonants of distorted words when these words had been produced by the same talker as the one heard at training than when the words had been produced by a different talker, even though, as demonstrated by our control experiments, participants were not better than chance at deciding whether the distorted test words originated from the same or from a different talker. It is thus difficult to attribute the speaker-match benefit to a response bias. Because the magnitude of the syllable-position match effect in Experiments 1 and 3 was similar as that of the speaker-match effect in Experiment 3, it is reasonable to assume that both types of effects reflect an increased ability to correctly identify the linguistic components of noise-vocoded utterances as their acoustic similarity with the trained utterances increased. This conclusion corroborates other studies where improvement over time with a novel form of speech reflects genuine learning, rather than improved task strategies (Dupoux & Green, 1997; Hervais-Adelman et al., 2008; Peelle & Wingfield, 2005).

As reviewed in the introduction, a number of studies have documented listeners' ability to rapidly adapt to a novel form of speech (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Dupoux & Green, 1997). The present study departs from this past research by systematically controlling the relationship between training and testing stimuli. This enables us to understand the process by which listeners learn to adapt to an unusual form of speech. Our results suggest a process by which listeners, based on their experience with distorted utterances, build knowledge that reflects this experience. Importantly, this knowledge should not be conceived as an ensemble of unanalyzed acoustic images, but as a structured database that can be successfully applied to instances that are not identical to those that served to build the knowledge. Thus, people can show evidence of generalizing their experience to different utterances, but the degree of generalization remains dependent on the nature of this experience. Under this view, as experience with this form of speech increases and the similarity space formed by the set of experience utterances expands, abstraction from the set of experienced utterances should increase. Although the present study did not test this prediction, its findings are consistent with a learning process where generalization emerges from and is constrained by past experience.

Our findings confirm and expand on prior results that variability in the training stimuli improved listeners' ability to adapt to a novel form of speech. For example, Greenspan et al. (1988) trained listeners to understand synthetic speech by exposing them either to a small list of words, each one repeated several times, or to a larger set of words with fewer repetitions of each word, and found greater generalization in the latter group than in the former. This finding is consistent with our conclusion that listeners are not equally successful at transferring their experience with a sound pattern to predict the sound patterns associated with other members of the same category, even when the acoustic-phonetic structure of the distorted speech is lawfully and systematically related to the acoustic-phonetic structure of their native language.

Although the syllable-position effect on consonant generalization observed in the present study may suggest that allophones (as opposed to phonemes) are the categories listeners generalized their experience over, the additional influence of whether or not the consonant had been heard flanked by the same vowel on consonant identification leads us to doubt this conclusion. In the same way that the syllable-position effect indicates that generalization did not equally apply to all members of the same phonemic category, the vowel-context effect indicates that generalization did not equally apply to all members of the same allophonic category. Thus, the categories listeners generalized over are best conceptualized as structures that emerge in the similarity space formed by the memory traces of the noise-vocoded utterances.

This view of structure and categories emerging from language use is increasingly being embraced by many scholars (e.g., Beckman & Edwards, 2000; Bybee, 2001; Nygaard et al., 1994; Pierrehumbert, 2006). In the domain of phonology acquisition, it is now widely accepted that phonological categories largely result from experience with the ambient language, rather than being fixed by our biological endowment. For instance, studies have documented how languages differ in the way phonological categories divide the same phonetic space (e.g., Cho & Ladefoged, 1999). Language users, from a very early age, show evidence of having learned this partitioning (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Beckman and Edwards (2000) argue that the phonological categories that children and adults may possess have arisen from exposure to the words and phrases of their ambient language. A number of studies have suggested that the forces contributing to this categorization include the distributional properties of ambient sounds (Guenther & Gjaja, 1996; Maye, Werker, & Gerken, 2002; Peperkamp, Le Calvez, Nadal, & Dupoux, 2006).

The present study suggests that the process by which adult listeners learn to adjust to the phonological structure of a dialect or artificially altered speech proceeds in a similar way as that of building phonological categories in one's native language by young language learners. Importantly, functional categories, such as phonemes, played a modest role: Listeners' knowledge of contextual variation in phonemic category structure, that is, their knowledge of how variants of the same phonemes relate to one another in acoustic space, had a limited influence on their ability to generalize their experience with one member of this category to other members. For instance, people were not able to accurately predict, based on their past experience with a consonant in a given syllable position, how the same consonant would sound in the opposite position. We do not take our results to demonstrate that the experience that listeners have had with their language's phonological structure plays no role in their perception and adaptation to noise-vocoded speech. Rather, our results highlight the fact that its influence is limited. We return to this point below.

Our conclusions may appear at odds with those offered by other studies on talker adaptation, where listeners' ability to generalize their experience with a novel form of speech to untrained words was interpreted as evidence that listeners can adjust phonemic categories (Davis et al., 2005; Maye, Aslin, & Tanenhaus, 2008). However, as demonstrated in the present study, generalization to untrained words does not necessary imply adjustment of an entire category. Many of these studies did not test the degree to which generalization extended beyond the contexts in which tokens of the relevant categories were experienced. For instance, Maye et al. (2008) exposed listeners to a synthesized rendition of a segment from the well-known story "The Wizard of Oz" where front vowels had been categorically lowered. Following this exposure, listeners were more likely to accept a nonsense string as an English word than they were prior to hearing the story if shifting its vowel in a comparable manner formed an existing word. Importantly for current purposes, this effect was observed on words people heard shifted in the story as well as on stimuli that had not been heard. This result led the authors to propose that listeners adjusted their interpretation of any instance of a vowel category. However, it may be that generalization was mainly observed for untrained words in which the critical vowel occurred in a context phonetically similar to those experienced at training. Because Maye et al. did not document the extent to which the vowels' contexts at testing resembled those at training, the scope of the observed generalization is difficult to establish.

Our investigation of listeners' ability to transfer their knowledge on the interpretation of noise-vocoded speech from one speaker to another led to an intriguing finding. In the two studies that manipulated whether the testing utterances had been produced by the same talker as at training or by a different talker, the degree to which listeners successfully generalized the knowledge acquired during training was independent of whether people were able to discriminate the two talkers. This corroborates a previous report by Yonan and Sommers (2000) showing that cross-talker specificity, and thus, limited generalization did not hinge on listeners' ability to discriminate or recognize talkers. The aspect of our data that is harder to explain concerns Experiment 4's cross-talker generalization concurrently to talker discrimination. This result, we propose, illustrates listeners' ability to adjust, or normalize, the talkers' utterances based on their perceived voice differences (e.g., by factoring out the vocal-tract influence on the rendition of speech segments). This interpretation does not question the existence of talker-specificity effects in the processing of clear speech–such effects have been repeatedly demonstrated. But these robust effects are nonetheless subtle compared to humans' capacity to understand speech from any new talker. Experiment 4's results may be mostly capturing listeners' cross-talker generalization, which is possible if talkers' voices can be discriminated despite the noise-vocoded distortion.

We now turn to the implications of this work to speech-perception research. Our results have shown that people do not generalize their experience to novel cases as well when these cases are acoustically dissimilar to their experience as they do when novel cases are similar. Assuming that the ability to generalize is directly constrained by category boundary, one interpretation of our findings is that different realizations of the same sounds are grouped by their perceptual, as opposed to functional, similarity. This is, in essence, the approach that proponents of an episodic view of the lexicon advocate for (see Goldinger, 1998). An alternative view, the one described in the introduction, is that function plays a major role in organizing or re-organizing listeners' perceptual categories. Thus, acoustically dissimilar sounds may be grouped together because of their functional equivalence. This view cannot be ruled out based on the present data. Indeed, the ability to correctly identify a new instance as a member of a category exemplified at training requires listeners to derive, based on its noise-vocoded rendition, its relationship with the instance that was previously experienced. Even though the distortion imposed by the noise-vocoded transformation is lawful, this computation is necessarily more complex for functional categories (i.e., categories that group instances that may be perceptually distant) than for perceptual categories. Thus, the absence of evidence for functional categories cannot be interpreted as evidence against their existence and role in speech perception because their emergence is necessarily more complex than that of perceptual categories.

The present study also examined the role that lexical knowledge may have on listeners' ability to adapt to a novel form of speech. Listeners were trained to understand noise-vocoded speech by listening to utterances that formed either existing English words or nonsense strings. In both cases, listeners received feedback in the form of the original, undistorted version of each utterance. As reviewed in the introduction, the recognition of sounds embedded in a word context is facilitated because the set of possible sounds is more constrained in a word than in a nonsense string. Experiment 2, wherein the frequency of segments and bigrams was perfectly matched between the two training sets, provided the most definitive test of this question, and the results are intriguing: Listeners were no more likely to correctly identify trained consonants after having been exposed to word stimuli than after having been exposed to nonword stimuli. Thus, there is little evidence that the word contexts provided revealed the identity of the consonants in training stimuli more effectively than nonword contexts did. However, an advantage for word training emerged when assessed on people's ability to correctly identify English consonants that had not been heard during training. This result indicates that people were better able to transfer their experience with noise-vocoded speech to novel stimuli when this experience had consisted of hearing familiar strings.

The fact that the training regime was found to affect generalization to untrained consonants provides some evidence that the sound patterns (i.e., the words and frequent phrases) that give rise to the phonological categories and structure of a language continue to exert their influence in listeners' speech perception (cf. Bybee, 2001). If phonological categories had entirely abstracted from these sound patterns, no difference in the training regimes would be expected because the two training sets were perfectly balanced in terms of these phonological categories (i.e., their frequency of occurrences and the local phonetic contexts in which their members occurred). Under this view, once a category has emerged, exemplars from this category would be equivalent with one another in all respects, regardless of the lexical context in which they occur. Our result indicates that people were better able to generalize after exposure to words than nonwords. Thus, we must conclude that listeners' past experience with words' sound patterns did contribute, albeit modestly, to their learning to interpret noise-vocoded speech. The challenge consists in developing a framework where word knowledge exerts its influence on listeners' generalization. We also acknowledge that our interpretation does not readily explain why no effect of the training regime (word vs. nonword) was observed on trained consonants. Further research is necessary to fully account for this aspect of the results.

To summarize, the present study examined the process by which people learn to comprehend artificially distorted speech. In each of the experiments presented here, people were trained to decode noise-vocoded speech by hearing CVC stimuli both in their distorted and unaltered forms. After a relatively brief exposure to noise-vocoded stimuli, listeners could generalize their experience with distorted stimuli to novel words. However, this ability did not result from a complete remapping of listeners' phonemic categories, as evidence of cross-word transfer is sometimes interpreted. Instead, the probability of accurately recognizing words or consonants depended on the phonetic properties of the testing words with respect to the training set. People recognized previously heard consonants better if they had been heard in a similar context (i.e., its syllable position and vocalic context) than in a dissimilar context. Varying the identity of the talkers between training and testing had also an influence on recognition, although not systematically. Thus, even though the noise-vocoded transformation involves a systematic and lawful distortion of the acoustic signal, we found that listeners' adaptation involved only a modest transfer to members of the phonological categories other than those experienced. We also found some evidence of better generalization when listeners had been trained on existing English stimuli than nonsense, unfamiliar strings. We take this as evidence supporting a view of phonology where categories and structure emerge from the ambient words and sound patterns without completely abstracting from them.

References

Baayen, R. H. (2008). Analyzing linguistic data. In *A practical introduction to statistics using R.* New York: Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459.

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development, 71,* 240–249.

Benkí, J. R. (2003). Quantitative evaluation of lexical status, word frequency, and neighbhorhood density as context effects in spoken word recognition. *The Journal of Acoustical Society of America, 113,* 1689–1705.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved from http://www.praat.org

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of Acoustical Society of America, 84,* 101–113.

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

Brownman, C. P., & Goldstein, L. M. (1995). Gestural syllable position effect in American English. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing speech: Contemporary issues for Katherine Safford Harris* (pp. 19—34). Woodbury, NY: American Institute of Physics Press.

Bybee, J. (2001). *Phonology and language use.* New York: Cambridge University Press.

Cho, T., & Ladefoged, P. (1999). Variations and universals in VOT: Evidence from 18 languages. *Journal of Phonetics, 27,* 207–229.

Christie, W. M. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America, 55,* 819–821.

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

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,* 222–241.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 914–927.

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

Francis, A. L., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception and Psychophysics, 62,* 1668–1680.

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effects of training on the acoustic-phonetic representation of synthetic speech. *Journal of Speech, Language, and Hearing Research, 50,* 1445–1465.

Fu, Q.-J., Chinchilla, S., Nogaki, G., & Galvin, J. J. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *Journal of Acoustical Society of America, 118,* 1711–1718.

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

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 421–433.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of Acoustical Society of America, 100,* 1111–1121.

Hervais-Adelman, A., David, 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.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Upper Saddle River, NY: Prentice Hall.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446.

Kouider, S., & Dupoux, E. (2005). Subliminal speech priming. *Psychological Science, 16,* 617–625.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review, 13,* 262—268.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 225,* 606–608.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica, 5,* 1–54.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68,* 1–13.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32,* 543–562.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infants sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113–1126.

Miller, G. A., Bruner, J. S., & Postman, L. J. (1954). Familiarity of letter sequences and tachistoscopic identification. *Journal of General Psychology, 50,* 129–139.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329–335.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America, 62,* 714–719.

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.

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1315–1330.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition, 101,* B31–B41.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics, 34,* 516–530.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–949.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395–408.

Shannon, C. W. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423, 623–656.

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

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics, 68,* 1–16.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory & Language, 48,* 233–254.

Sproat, R., & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics, 21,* 291–311.

Sternberg, S., Wright, C. E., Knoll, R. L., & Monsell, S. (1980). Motor programs in rapid speech: Additional evidence. In R. A. Cole (Ed.), *Perception and production of fluent speech (Fourteenth Annual Carnegie Symposium on Cognition)* (pp. 507–534). Hillsdale, NY, Erlbaum.

Warren, P., & Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics, 41,* 262–275.

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics, 50,* 351–360.

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging, 15,* 88–99.

Appendix A

Stimuli Used in Experiment 1

Testing word	Onset-nucleus-rime overlap		Onset-coda overlap		Consonant overlap		Word	
				Training stimuli				
bane	beIz	teIn	bɔv	θɔn	vɔb	nɔs	knob	name
bathe	beIp	tʃeIθ	bɛp	tʊð	kIb	ðIl	cub	this
boys	bɔIð	fɔIz	baIʃ	mæz	ʃɛb	zɔt	rib	zeal
check	tʃɛd	sɛk	tʃin	loʊk	ðʊtʃ	kɔθ	peach	couch
chive	tʃaIɹ	taIv	tʃæz	gæv	nitʃ	vum	ditch	vase
far	fɑz	lɑɹ	foʊd	tʃɔIɹ	reIf	roʊs	reef	rug
fudge	fʌv	lʌdʒ	fɔð	tɔdʒ	dɔf	jɔt	puff	job
gap	gæv	θæp	gʊdʒ	mʊp	vʊg	pʊtʃ	dog	path
good	gʊð	mʊd	gIdʒ	tʃId	ðɛg	dɛtʃ	shag	deaf
gouge	gaʊd	baʊdʒ	gɑɹ	lIdʒ	dig	dʒaIs	jog	jail
lewd	luʃ	gud	lʊdʒ	sɔId	nɔIl	dɔIg	veal	doubt
look	lʊp	tʃʊk	leIv	θik	ɹæl	kaʊt	coil	coach
mop	mɔn	fɔp	meIʃ	beIp	jaʊm	paʊf	shame	put
mush	mʌn	sʊʃ	miθ	biʃ	pum	ʃub	them	shawl
seed	siʃ	θid	suk	fɔd	ʃoʊs	doʊg	voice	pig
soar	sɔv	θɔɹ	saIʃ	θaIɹ	kaIs	raIθ	juice	roof
teethe	tiʃ	miθ	tɔIp	fuð	pɔIt	ðɔIm	that	thus
thighs	θaIk	tʃaIz	θoʊn	suz	ʃæθ	zæf	death	zoom
thin	θIdʒ	gIn	θoʊk	goʊn	vuθ	næm	wrath	night
toes	toʊdʒ	moʊz	taʊk	faʊz	dʒeIt	zeIb	shout	zing

Appendix B

Training Stimuli Used in Experiment 2

Word	Nonword	Word	Nonword
boat	bok	folk	fot
book	bʊt	foot	fʊk
dame	deIk	fuss	fʌm
debt	dɛm	gem	dʒɛk
debt[*]	dɛm[*]	germ	dʒɜ□s
deck	dɛm[*]	germ	dʒɜ□t
deck[*]	dɛs	jake	dʒeIm
deck[*]	dɛs[*]	jess	dʒɛk[*]
deem	dit	joke	dʒos
deem[*]	dit[*]	joke	dʒoʊm
dim	dIt	juice	dʒum
dime	daIt	jute	dʒum[*]
dime[*]	daIt[*]	hearse	hɜ□k
dirt	dɜ□m	hearse[*]	hɜ□m
dirt[*]	dɜ□m[*]	heat	him
dome	dok	heck	hɛt
doom	dut	height	haIm
dose	dok[*]	hem	hɛt[*]
dose[*]	dok[*]	hum	hʌs
feet	fim	hurt	hɜ□m
femme	fɛt	whom	hus
fight	faIm	let	lɛk
firm	fɜ□s	lurk	lɜ□t
firm[*]	fɜ□t	yes	jɛk
fit	fIm	yolk	jos

[*] Extra occurrences of the same training stimulus.

(*Appendices continue*)

Testing Stimuli

	List A		List B	
	Same position	Different position	Same position	Different position
	Testing word stimuli with trained consonants			
d	*dawn*			*loud*
	dull			*hide*
dʒ		*lodge*	*join*	
		wage	*jeep*	
k	*hook*			*cape*
	pick			*cop*
m		*mill*	*ham*	
		mall	*lamb*	
f	*fail*			*knife*
	fan			*laugh*
b		*lab*	*bang*	
		pub	*boil*	
t	*hate*			*tap*
	nut			*tune*
s		*sane*	*lace*	
		soap	*niece*	
	Testing word stimuli with untrained consonants			
g	*goal*	*gown*	*league*	*pig*
r	*ring*	*rope*	*liar*	*peer*
θ	*thin*	*thong*	*heath*	*path*
ʧ	*chain*	*chip*	*pitch*	*hatch*
ʃ	*shell*	*shine*	*wash*	*push*
v	*veil*	*vine*	*wave*	*nerve*
z	*zen*	*zap*	*wise*	*haze*

Appendix C

Training Stimuli Used in Experiment 3

Set A	Set B
dab	bad
cab	back
mob	bomb
rob	bar
rife	fire
calf	fad
deaf	fed
reef	fear
mag	gap
vague	gave
rogue	gore
mug	gum
vile	live
pal	lamb
dill	lid
veal	leave
Dan	knack
main	name
vain	knave
pin	nip
pat	tap
kit	tick
pit	tip
cut	tuck

Testing Stimuli

	Speaker RM		Speaker *SD*	
	Onset	Coda	Onset	Coda
b	bathe	chub	bass	sub
	beach	hub	booth	web
f	Fish	huff	phase	chafe
	Fuss	hoof	fudge	safe
l	Lace	jail	ledge	howl
	lodge	yell	loose	hole
t	tooth	height	teach	wet
	toss	hot	tease	sheet
n	noise	hen	nice	join
	nose	sign	notch	wine
g	goose	jig	guess	wig
	gouge	jog	gong	hog
d	daze	hood	dodge	seed
	dice	hide	dose	wade
k	catch	joke	cage	hike
	coach	sock	cause	wake
m	mouth	gem	mice	seem
	myth	home	miss	theme
p	page	soap	pace	weep
	pose	sheep	push	shape
r	race	chair	rage	wear
	wretch	hair	rush	sure
v	voice	shove	verge	have
	vows	wove	verse	sieve

Appendix D

Training Stimuli Used in Experiment 4

Set A (back vowels)	Set B (front vowels)
rude	read
roam	rim
rod	raid
rum	ram
boom	beam
bowl	bid
bomb	bake
bud	bad
poke	peak
pool	pill
pall	pail
puck	pal
nuke	nick
node	name
knock	knack
null	kneel

(*Appendices continue*)

Testing Stimuli

	Speaker RM	Speaker CN		Speaker RM	Speaker CN
onset			coda		
b	batch	base	d	fad	fade
	beef	bit		hid	food
	booth	botch		thud	god
		both		toad	seed
		bus	k	juke	chuck
n	gnash	Nate		sock	tack
	knit	niece		thick	weak
	knot	nudge		wake	yolk
	noose		l	heel	chill
	nose			sale	full
p	pave	patch		soul	gal
	peach	pig		wall	tool
	poach	pooch	m	gum	home
	puff	pot		ham	shame
r	rave	roof		theme	Tom
	ridge	rose		zoom	whim
	rot	rag			
	rush	wreath			

Received July 17, 2008
Revision received June 16, 2009
Accepted June 23, 2009 ■