LANGUAGE AND COGNITIVE PROCESSES 2011, iFirst, 1–18

Psychology Press
Taylor & Francis Group

Word learning under adverse listening conditions: Context-specific recognition

Sarah C. Creel[1], Richard N. Aslin[2], and Michael K. Tanenhaus[2]

[1]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA
[2]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

Previous studies of word learning have presented the items to listeners under ideal conditions. Here we ask how listeners learn new vocabulary items under adverse listening conditions. Would listeners form acoustically-specific representations that incorporated the noise, base their representations on noise-free language knowledge, or both? To address these questions, listeners learned 16 words as labels for unfamiliar shapes presented on a computer display. During the learning phase, word-shape pairings were presented with either clear or white-noise-embedded tokens. For each word (e.g. *dabo*), another word shared consonants (e.g. *dubei*) and a third shared vowels (e.g. *gapo*). Learning was assessed in a 4AFC picture-selection task. The highest accuracy and speed were achieved by listeners who experienced the same noise level at exposure and test (both clear or both noisy), suggesting that listeners' representations of noisy words were faithful to the spectral context experienced during the learning phase. Implications for word learning and recognition across a variety of listening conditions are discussed.

Keywords: Word learning; Cue weighting; Speech in noise; Representational specificity.

A typical speaker's vocabulary consists of more than 50,000 words, and many words differ from each other only by subtle acoustic features (e.g. "cat" vs. "pat"). Thus, learning a new word requires activating a sophisticated system of detecting and remembering fine-grained phonetic information. At the same time, individual exemplars of the same word can differ drastically across contexts, such as talker (male vs. female voice), speaking rate, dialect, and ambient noise. The learner's challenge, therefore, is to determine which perceived acoustic elements matter for identifying words, so that the listener can selectively attend to those lexically relevant attributes without being misled by other attributes.

Complicating matters further, even the "meaningful" elements in a word (/k/ vs. /p/ in *cat* vs. *pat*) may be more or less available from one listening situation to another.

Correspondence should be addressed to Sarah C. Creel, Department of Cognitive Science, University of California, San Diego, Mail Code 0515, 9500 Gilman Drive, La Jolla CA 92093-0515, USA. E-mail: creel@cogsci.ucsd.edu

This work was supported by NIH grant DC 005071 to MKT and RNA, a Packard grant 2001-17783 to RNA, and an NSF graduate research fellowship to SCC.

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2011.610597

How do listeners recognise speech when recognition cues are only probabilistically available? A substantial area of interest in language learning and perceptual adaptation is how listeners come to recognise words under poor listening conditions. Do listeners form completely novel representations for recognition in adverse situations (having a representation of clear "cat" and another for noisy "cat"), or do they use the same representations of a word across different listening situations? On the one hand, recognition would operate best if the listener's representations matched the listening context, suggesting that specific representations (word + context) are more useful. On the other hand, the listener also needs to be able to recognise the same word across different speakers and listening conditions, implying that use of the same representation, perhaps with adjustments for context according to the listening situation, would be beneficial.

The goal of the current work was twofold. First, we wished to understand the level of specificity of the lexical representations underlying listeners' recognition in adverse listening situations. Second, we wanted to understand whether learning in altered listening situations is dependent on flexible adjustment of listeners' existing weightings of different cues to word identity. To motivate these questions, we first discuss what is known about the specificity of listeners' word representations. We then explore how these weightings might be flexibly adjusted in word recognition.

SPECIFICITY OF REPRESENTATIONS

A major "unknown" in understanding processing of altered speech is how listeners represent an unusual variant of a known word. For instance, does an American English speaker represent an r-less British pronunciation of "father" as a variant of the American r-ful version, or does the American speaker maintain two separate representations? Work by Sumner and Samuel (2009) suggests that listeners who learn a second accent later in life preferentially encode "canonical" forms (those of their native accent), suggesting that representations of the new accent may be encoded as alterations to the originally learned one, with the original form serving as an anchor point. Interestingly, more balanced exposure during an early period of word learning seems to lead to dual representations (Sumner & Samuel, 2009). Studies of perceptual adaptation, which limit the kinds of changes contained in a natural dialect to a single variable, demonstrate that the degree of shift from familiar forms—such as the size of an upward shift in frequency of noise-vocoded speech—affects the ease of adaptation (e.g. Rosen, Faulkner, & Wilkinson, 1999). These results, like those for dialect shifts, suggest that listeners base their learning of altered speech input on familiar forms. Under adverse listening conditions, therefore, listeners might also represent altered speech by adapting their preexisting representations, rather than forming new ones.

An alternative to the foregoing "prototype + adaptation" model is that listeners form acoustically-specific representations of words in adverse listening situations. This is consistent with recent work suggesting that acoustic detail is particularly important in situations of energetic masking (ambient noise; Mattys, Bradlow, Davis, & Scott, 2011 this issue; Mattys, Brooks, & Cooke, 2009). Recent work on acoustic specificity in word recognition suggests that listeners store acoustic details of a talker's voice (Goldinger, 1996, 1998; Palmeri, Goldinger, & Pisoni, 1993), either as an abstract representation of the talker's phonology (Cutler, Eisner, McQueen, & Norris, 2010; Eisner & McQueen, 2005), as part of the word itself (Creel, Aslin, &

Tanenhaus, 2008), or both (Jesse, McQueen, & Page, 2007). Studies of accent adaptation suggest that adaptation to a particular accent does not necessarily generalise to an unfamiliar accent (Bradlow & Bent, 2008) or talker (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), implying that acoustically-specific properties may facilitate accented speech recognition. Extending this "context-specificity" model to adverse listening contexts predicts that listeners would show better recognition of materials learned under adverse conditions than those learned initially under good (i.e. noise-free) conditions.

So far we have considered how the prototype + adaptation and context-specificity models operate on preexisting word representations. An additional challenge is how listeners deal with altered speech input when they do not know the corresponding canonical form—for instance, when confronted with new words in the presence of some acoustic distortion? In addition to understanding the effects of *perception* in adverse listening conditions, it is important to understand the effects of *learning* in adverse conditions. Word learning continues throughout life (e.g. "staycation" and "truthiness"), under a variety of listening conditions, and even as older listeners begin to experience diminished hearing capacities. Are listeners better at recognising words in adverse listening conditions when they have initially formed canonical representations of words, or when they have formed specific representations which encompass the distortion from a prototype created by noise?

CUE WEIGHTING ASYMMETRIES

If and when listeners utilise existing representations in recognising an impoverished or distorted speech signal, how do they do so? The most obvious solution would be adjusting their weighting of existing cues to speech sounds, so that cues less affected by the distortion are given greater weight in recognition. Work by Mattys (2004) supports the idea that cue weightings may vary flexibly with the listening situation. In Mattys' study, listeners were asked to segment (place word boundaries in) a short string of syllables. Mattys found that coarticulatory cues influence word segmentation more than stress placement in clear listening conditions, but stress has a stronger influence in noisy conditions (see Miller & Wayland, 1993, for a similar pattern of results in /b/-/w/ discrimination). Further work suggests that listeners exposed to accented speech flexibly shift phoneme boundaries (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006; Maye, Aslin, & Tanenhaus, 2008) and that listeners change the weightings of cues when the distributional properties of the categories are altered (Clayards, Tanenhaus, Aslin, & Jacobs, 2008).

Given that listeners can flexibly re-weight acoustic cues to adjust to many speech contexts, do they deploy a similar mechanism when listening to speech under adverse conditions? Listeners can learn to recognise highly impoverished speech representations that mimic many of the distortions present under noisy listening conditions. For example, after a short period of training, listeners can recognise words in sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981), time-reversed speech (Saberi & Perrott, 1999), and noise-vocoded speech (which mimics the sound input of individuals with cochlear implants; Dahan & Mead, 2010; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Shannon, Zeng, Kamath, Wygonsky, & Ekelid, 1995). The ability to rapidly and flexibly re-weight cues in a context-specific manner would be helpful in a

world where listening conditions vary from moment to moment, and where optimal listening conditions may not be modal.

THE CURRENT STUDY

The purpose of the present study is to assess the *specificity* of listeners' representations of altered speech, and the *flexibility* of their cue weightings in unsupervised learning situations. To assess specificity, listeners learned words in the clear or in noise, and were tested in either the same listening conditions or the changed listening conditions. To assess cue re-weighting, we examined the effect of noise on consonant–vowel weighting, a cue asymmetry that has received recent attention in the literature. Across several different paradigms, languages, and age groups, listeners seem to give more weight to consonants than to vowels in word recognition (Bonatti, Peña, Nespor, & Mehler, 2005; Creel, Aslin, & Tanenhaus, 2006; Cutler, Sebastián-Gallés, Soler-Vilageliu, & Van Ooijen, 2000; Nazzi, 2005; Nespor, Peña, & Mehler, 2003; Van Ooijen, 1996; though see Newport & Aslin, 2004), possibly because consonants are more informative for lexical identity than vowels (Altmann & Carter, 1989). For instance, Creel et al. (2006) found that listeners confused newly learned words like *bamo* with consonant-matched words (*beimi*) more than with vowel-matched words (*gapo*). Collectively, these results suggest that listeners use consonants more to identify words than they use vowels.[1] Might listeners adapt to noisy learning conditions by giving different weightings to existing representations of vowels vs. consonants?

To explore these issues, listeners learned an artificial vocabulary, similar to that of Creel et al. (2006), in clear or noisy listening conditions. Frozen white noise was used because it affects consonant perception more than vowel perception (Horii, House, & Hughes, 1971). The word "frozen" indicates that the noise pattern for a given word was consistent across all presentations of that word, which mimics a consistent external-noise distortion of the input. For each of the 16 words in the vocabulary (e.g. *dabo*; see Appendix 1), another word matched its consonants (*dubei*), and a third matched its vowels (*gapo*). Each word was presented as a label for a unique and initially unfamiliar black object on a white background. Following this exposure phase, participants were tested on their ability to identify words with or without noise.

We wanted to know whether listeners would specifically encode the distortion along with the newly learned words. If so, they should recognise words better when distortion was maintained from learning to test. Alternatively, listeners might benefit from learning under familiar (noise-free) conditions, suggesting that canonical representations are more effective for recognising words under acoustic alteration. Related to canonical representations, we wanted to know whether listeners would adapt to noise distortion by adjusting their existing weightings of known phonological categories—consonants vs. vowels—in recognition. If so, listeners should show an increase in vowel confusions relative to consonant confusions when words are learned in noise—a pattern which would indicate that vowel information was up-weighted under these conditions.

[1] As an important aside, these consonant-vowel differences in cue weighting may apply primarily to syllable-onset consonants rather than coda consonants. Coda consonants are rarer crosslinguistically, and more subject to misidentification (Redford & Diehl, 1999) and recognition failure (Creel & Dahan, 2010). Further, Creel et al. (2006) found that words sharing vowels were more confusable than words sharing coda consonants.

METHOD

Participants

Participants were 92 University of Rochester undergraduates who did not report a history of hearing problems. They were paid $10 each for an experiment that lasted 20–30 min.

Stimuli

Words

There were 16 consonant-vowel-consonant-vowel (CVCV) words in the artificial lexicon, constructed from the consonants b, d, g, and p, and the vowels a, e, i, o, and u. The consonants were selected such that three differed only by a place of articulation contrast. Consonantal place is known to be especially vulnerable to noise, such as multi-talker babble (Cutler, Weber, Smits, & Cooper, 2004) and noise-band vocoding (Shannon et al., 1995). Each word had one other word that shared both its consonants, and one other word that shared both its vowels. Words were generated using the MacInTalk speech synthesiser, voice Victoria, and the SpeechSaver utility (Singer & D'Oliveiro, 2001), at a sampling rate of 44100 Hz. Using a synthetic voice assured uniform amplitude and production throughout all recorded words. Words plus frozen noise samples were created by adding to each sound file a randomly-generated white noise vector of the same length as the word, yielding a signal-to-noise ratio of 6.7 dB. This is illustrated in Figure 1a (clear) and b (noise-embedded). Files were saved in SoundEdit16 format for presentation in PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993).

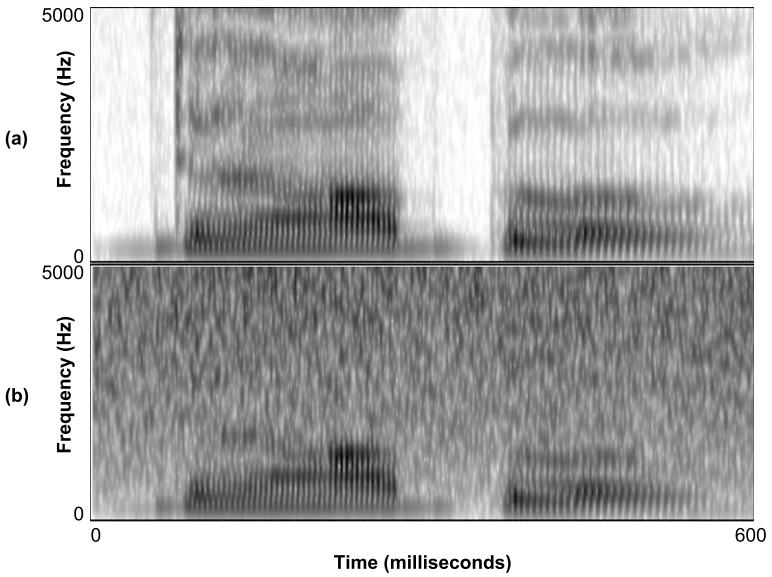

Figure 1. Spectrogram of novel word (*dabo*) in the clear (a) and embedded in white noise (b).

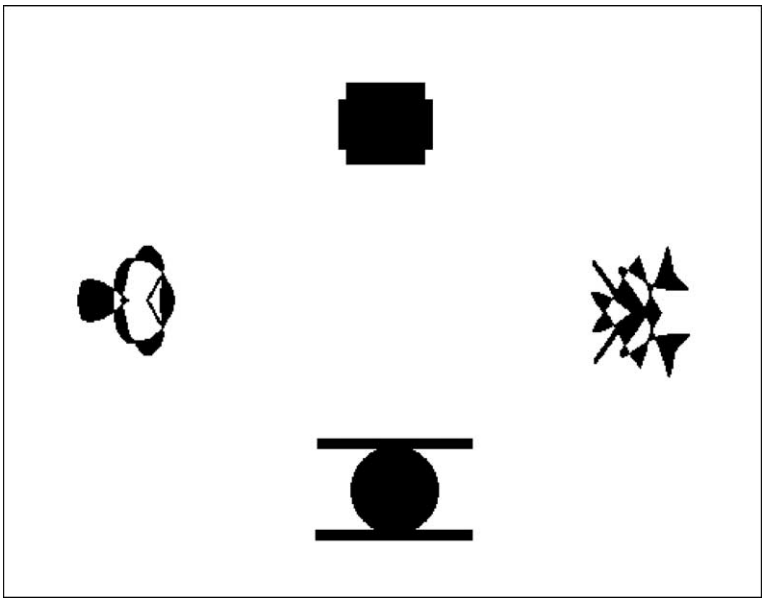

Figure 2. Test display. On each trial, participants selected one of four pictures as the target when a word was spoken.

Pictures

The 16 words were used as labels for a set of 16 unfamiliar drawings (Figure 2), originally created in AppleWorks Paint and used in several earlier studies (Creel et al., 2006, 2008; Creel & Dahan, 2010; Creel, Tanenhaus, & Aslin, 2006).

Counterbalancing

To minimise the effects of particular words being idiosyncratically easy to learn, there were four different random assignments of consonants and vowels to syllable positions within words (Appendix 1), each heard by a different set of participants. Crossed with this, there were six different assignments of pictures to labels, minimising the possibility that particular word-picture pairings would be more easily encoded than others. This generated 24 unique exposure lists, each of which could occur in each noise condition. Thus, a particular picture could be labeled 24 different ways across participants within a noise condition. Nearly all lists were presented in the all-clear, all-noise, clear-to-noise, and noise-to-clear conditions. The all-noise condition was completed (all 24 lists). The noise-to-clear condition was almost completed except that one list was run twice while another was left out. All-clear and clear-noise conditions were not quite completed (22 out of 24).

Procedure

Exposure

Listeners were exposed to words in noise or in the clear, and were tested on words in noise or in the clear. Both exposure noise levels were crossed with both test noise levels, yielding four between-participants conditions: clear exposure–clear test (“all-clear,” $n = 22$), clear exposure–noise test (“clear-to-noise,” $n = 22$), noise exposure–clear test (“noise-to-clear,” $n = 24$), and noise exposure–noise test (“all-noise,” $n = 24$).

PsyScope experimental presentation software (Cohen et al., 1993) was used to present the stimuli. On each exposure trial, a picture appeared in the center of the screen, and after 750 ms, its name was spoken. The participant then mouse-clicked the picture to proceed to the next trial. Each word and its paired picture were presented 24 times, in the center of the screen (200 × 200 pixels), for a total of 384 exposure trials.

Test

In testing, four pictures at a time appeared: above, below, to the left, and to the right of center, as depicted in Figure 2. After 100 ms, one of the pictures was named, and then the word "Next" appeared in the center of the screen. Participants clicked on the picture that they thought had been named. The mouse-click caused the four pictures to disappear and participants clicked on the word "Next" to proceed to the next test trial. Each picture appeared four times as a target: twice with its same-consonant competitor's shape present, and twice with its same-vowel competitor's shape present, for a total of 64 test trials (32 trials repeated once each; see Appendix 2). The other two pictures in a trial were phonologically unrelated to the target, overlapping in none of the four segment positions (CVCV).

Processing of data

The chi-squared criterion for nonchance performance in a 64-item four-alternative forced-choice task is 35.94% correct ($p < .05$). Seven participants did not meet this criterion and were eliminated from analyses (all-clear: 0, 22 participants remaining; all-noise: 1, 23 participants remaining; noise-clear: 4, 20 participants remaining; clear-noise: 2, 20 participants remaining). One more participant failed to respond on a large proportion of trials and was eliminated from the analyses (all-noise condition, 22 participants remaining). Thus, the final sample consisted of 84 participants. Data were analyzed using mixed-effects models that treat both participants and items as random effects.

RESULTS

We first evaluate the specificity of listener's representations by examining the overall accuracy and response time data. Then, to evaluate changes in cue weighting for vowels vs. consonants, we examine the particular types of errors (choosing consonant-match competitors or vowel-match competitors) that participants made under different learning and testing conditions.

Accuracy

Here we asked whether listeners form *specific* representations of altered input. If so, then accuracy should be greater for participants who experienced the same listening conditions during learning and test, regardless of whether these conditions were clear or noisy. An alternative prediction might be that accuracy would be greater for canonical exposure—that is, after exposure in clear listening conditions. Based on this, participants who learned in clear listening conditions, regardless of test conditions, should be more accurate. Overall accuracy suggested that listeners do form specific representations. That is, participants were more accurate (Figure 3) when they learned and were tested under the same conditions—clear *or* noisy—than when conditions

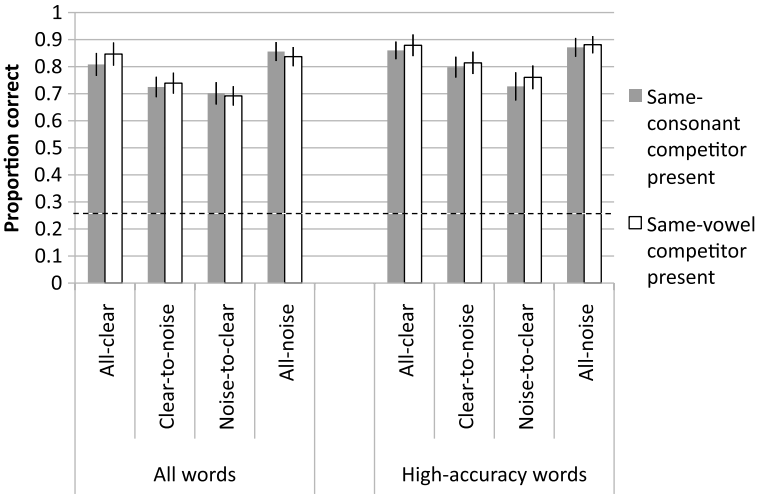

Figure 3. Overall accuracy (mean ± standard error) in each condition. Left side: all words; right side: words that were identified most accurately in noise. Dashed line corresponds to chance performance.

changed from exposure to test. Response times (Figure 4) also supported this hypothesis.

A logistic mixed-effects model was used to analyze the accuracy data in R (2008), using the Design package (Harrell, 2009) and the languageR package (Baayen, 2010). This analysis assumes, unlike ANOVA, that responses are distributed binomially, which better accounts for response variance. Exposure Noise (clear, noisy) and Test Noise (clear, noisy) were between-participants factors, and Segment (consonant competitor present, vowel competitor present) was a within-participants factor. For all analyses reported, all random effects of participants and items were tested for significance, but only those that significantly increased the variance accounted for were included in the final model. In all cases, the participants intercept term was included by convention.

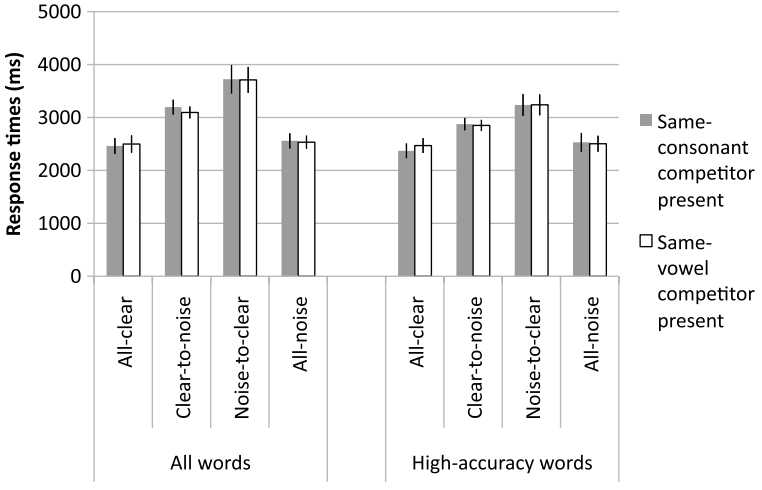

Figure 4. Response times (mean ± standard error) in each condition. Left side: all words; right side: words that were identified most accurately in noise.

To correct for the fact that the chance rate of correct responses was 1 in 4 rather than 1 in 2, the binomial distribution was offset by adding a logit(1/4) offset variable to our data frame in R, which specified the null hypothesis that 1/4 of the responses would be correct. The intercept of the model reached significance (coeff. = 3.08, $z = 16.10$, $p < .0001$), indicating that listeners responded correctly more often than chance would predict. No main effects reached significance, including the effect of Exposure Noise (coeff. = .067, $z = 0.37$, $p = .71$). This provides little support for an advantage in learning canonical representations. Only two interactions reached significance. One was Exposure Noise × Test Noise (coeff. = .674, $z = 3.77$, $p = .0002$), with better performance when the noise level was matched across exposure and test than when it changed. There was also an interaction of Exposure Noise × Segment (coeff. = .096, $z = 2.46$, $p = .01$), resulting from a small advantage on vowel-competitor trials for listeners exposed in the clear, with the reverse (higher accuracy on consonant-competitor trials) for noise-exposure conditions. Comparing individual exposure and testing conditions, all-clear participants were more accurate than both noise-to-clear (coeff. = .88, $z = 3.10$, $p = .002$) and clear-to-noise (coeff. = .66, $z = 2.40$, $p = .02$). Similarly, all-noise participants were more accurate than noise-to-clear (coeff. = .70, $z = 3.01$, $p = .003$) and clear-to-noise (coeff. = .60, $z = 2.52$, $p = .01$). No other paired contrasts were significant. Overall, the results suggest that a match between acoustic conditions during exposure and testing, but not the clarity of initial exposure conditions, is important for later recognition, supporting the specificity hypothesis.

Response times

Inspection of Figure 4 suggests that listeners in conditions where stimuli did not change from exposure to test performed equally well, regardless of whether stimuli were clear or in noise. It remains possible, though, that all-noise participants were only performing well because they were exerting substantially more effort than all-clear participants. One index of effort is response times. Therefore, we examined response times on correct trials in each condition in a mixed-effects model predicting response times with Exposure Noise (between-participants), Test Noise (between-participants), and Segment (within-participants) as factors. Response times for each participant that fell more than two standard deviations outside that participant's mean were eliminated. *p*-Values for this analysis are not provided by R, so they were derived from the normal approximation to the *t*-distribution; the anti-conservativity of this approximation is minimal when the number of observations is much greater than the number of parameters (Levy, personal communication).

The task was not a speeded one, so results should be interpreted with caution. Nonetheless, this analysis clearly demonstrated that participants who did not experience a change in listening conditions responded faster than participants who did experience a change in listening conditions (Exposure Noise × Test Noise, *estimate* = 500.4, $SE = 100.9$, $t = 4.96$, $p < .0001$); no other main effects or interactions reached significance. Note that this interaction term would be significant even for a *t*-test with a single degree of freedom—the most stringent test possible. Individually, the no-change conditions were each faster than change conditions (all-clear vs. clear-to-noise: *estimate* = 359, $SE = 111$, $t = 3.23$, $p = .001$; all-clear vs. noise-to-clear: *estimate* = 482, $SE = 93$, $t = 5.16$, $p < .0001$; all-noise vs. clear-to-noise: *estimate* = 315, $SE = 108$, $t = 2.92$, $p = .004$; all-noise vs. noise-to-clear: *estimate* = 632, $SE = 165$, $t = 3.82$, $p = .0001$). Clear-to-noise participants were faster than noise-to-clear

participants (*estimate* = 320, *SE* = 159, *t* = 2.01, *p* = .04), consistent with a small advantage for exposure to canonical (clear) representations during learning. Participants in no-change conditions (all-clear or all-noise) responded with equivalent quickness (*estimate* = 45, *SE* = 109, *t* = 0.41, *p* = .68), suggesting that all-noise participants did not expend vastly more effort in recognition than all-clear participants.suggesting that all-noise participants did not expend vastly more effort in recognition than all-clear participants.

In sum, listeners were strongly affected by the specificity but not the clarity of initial learning conditions. They were faster and more accurate when exposure and testing conditions matched, even when conditions were noisy. They were slightly faster, but not more accurate, when exposure conditions were clear.

There is a potentially problematic issue with the specificity hypothesis. Specifically, what if the effect of noise was to alter words to consistently sound like something else—that is, what if listeners perceived *dabo* in the clear as *dabo*, but in noise it was consistently perceived as *dago*?[2] This would make "switched" conditions harder because the words themselves would be perceived as phonemically different, rather than different at a finer grain of acoustic specificity. If each noise-embedded word is *consistently* perceived as something other than the clear word, then it is not surprising that listeners would misidentify it when the word was changed.

Recall that the noise-embedding was designed to make perception of the words' segments noisier, thereby increasing uncertainty. However, it was designed not to *distort* perception. That is, we assumed that the modal perception of each segment would be the same as the original. To verify this assumption, we conducted a transcription task on both the clear tokens and the noise tokens ($n = 13$ participants). A few trials were missing from one participant; thus, there were at least 12 transcriptions of each of the 64 words (4 lists × 16 words each). Certainty ratings of transcriptions were also obtained. The results, described below, confirmed that listeners are relatively accurate at identifying noisy segments.

One interesting aspect of the data was that listeners reported glides or liquids in some of the words, in similar proportions for both clear tokens ($M = 30.3\%$, $SD = 30.3\%$) and noise tokens ($M = 31.8\%$, $SD = 27.3\%$; not significant, $p = .57$). These insertion errors, such as reporting "blogu" for *bogu*, seemed to be related to the diphone synthesiser, which created a slight auditory discontinuity mid-syllabically because the two halves of the vowel are imperfectly matched. Because these responses were so prevalent, we counted them as correct responses.

Accuracy was lower overall for words in noise ($M = 47\%$, $SD = 23\%$) relative to clear words ($M = 85.3\%$, $SD = 16\%$; $t(63) = 11.74$, $p < .0001$). However, accuracy in identifying each segment in noise in each position was 70% or higher (first consonant, $70 \pm 23\%$; first vowel, $81 \pm 21\%$; second consonant, $83 \pm 19\%$; second vowel, $86 \pm 17\%$), suggesting that in a majority of cases, listeners identified the segments in the word accurately (i.e. their predominant response was the same in both listening conditions). This is what one would expect if the noise causes perception to be less certain (increasing variability without changing the modal perception). In addition, certainty ratings were lower for words in noise than for their clear counterparts ($t(63) = 18.53$, $p < .0001$). This is also consistent with the hypothesis that noise made listeners less certain of their word identifications.

Nonetheless, it is still possible that some subset of individual words might have been perceived qualitatively differently in noise vs. in the clear, and that these words

[2]We thank Arthur Samuel for pointing out this alternative.

might be carrying the effects at hand. To address whether this was the case, we computed, for each word, the proportion of correct recognitions for each segment. We then found the words for which each segment was reported accurately at least 67% of the time ($M = 88\%$, $SD = 5\%$, compared to $95 \pm 3\%$ for the same words in the clear)—that is, each segment was identified as the correct sound at least twice as often as anything else. We included cases where listeners reported an inserted liquid or glide, because such reports were so prevalent even in the clear tokens. The selected tokens did not differ in how many had inserted segments in the clear tokens (21.4%) versus the noise tokens (22.5%; $t(27) = 0.33$, $p = .75$). Twenty-eight of the 64 words fit this stringent criterion. (A similar set of words resulted if we restricted selection to reports that did not contain inserted segments.)

The main analyses on accuracy and response time were re-run on this subset of the test items. For accuracy, the interaction of Exposure Noise × Test Noise was still significant (coeff. = .53, $z = 2.40$, $p = .02$), suggesting that the clear exposure–clear test and noise exposure–noise test still exceeded the changed conditions in accuracy. The pattern of results for highly-accurate words (Figure 3, right side) is qualitatively similar to the full dataset. We also reran the response time analyses with this subset of words and found that the Exposure Noise × Test Noise interaction was still significant (*estimate* = 316, $SE = 94$, $t = 3.36$, $p = .0008$; see right side of Figure 4), suggesting that *speed* of responding was still faster for all-clear and all-noise conditions, relative to noise-to-clear and clear-to-noise conditions. In both analyses, main effects did not approach significance ($z, t \leq 1.54$, $p \geq .12$). These results suggest that, even when limiting consideration to the words that were most accurately identified in noise, listeners nonetheless showed specificity effects in their accuracy and response times.

Thus, it seems that listeners learn white noise—not, in itself, a linguistic property—as well as they learn actual phoneme strings. This is concordant with recent work by Pufahl and Samuel (2010), in which words were recognised better when presented with the same background sound (e.g. a telephone) as on a previous presentation. The current results complement and extend those results to suggest that listeners encode even incoherent sound properties (white noise) along with words, in addition to coherent environmental sounds.

Consonant vs. vowel confusions

Next, we examined listeners' error patterns (Figure 5a and b) in the all-clear vs. all-noise conditions to determine whether listeners had adapted to poor listening conditions by re-weighting existing sound categories: consonants vs. vowels. Recall that we hypothesised that listeners might increase their weighting of vowel information relative to consonant information during learning in noise, rather than holding on to the disadvantageous consonant-biased weighting that is evident in studies conducted in good listening conditions (reviewed in the Introduction). This would predict a higher rate of same-vowel confusions than is seen in clear listening conditions.

To assess weightings of consonant vs. vowel information, we looked at the rates of errors to same-consonant competitors (reflecting reliance on consonants) and rates of errors to same-vowel competitors (reflecting reliance on vowels) in a logistic mixed model. These errors were compared to a baseline of unrelated-word errors, that is, selection of one of the other two incorrect alternatives present on a given trial. To correct for the fact that there were twice as many distractors in a trial (2) as competitors (1), the binomial distribution was offset so that the null hypothesis was a 1:2 ratio of errors.

We used a mixed logistic model of errors with Condition (all-clear, all-noise) as a between-participants factor and Segment (competitor, distractor) as a within-participants factor. The intercept was significant (coeff. = .37, $z = 2.74$, $p = .006$), reflecting more competitor errors than chance would predict. Condition also reached significance (coeff. = .28, $z = 2.60$, $p = .009$), suggesting that all-noise listeners' errors had a higher proportion of competitor errors than did all-clear listeners' errors. No other main effect or interaction reached significance. The absence of the interaction pattern limits the conclusions we can draw from this comparison of consonant and vowel competitors. If there is re-weighting of consonant vs. vowel information when listening context changes it is subtle enough that we did not reliably detect it. Note that in a model with all four conditions included, consonant similarity did *not* outweigh vowel similarity (compare Figure 5a and b): the intercept was significant (coeff. = .38, $z = 3.94$, $p < .0001$), indicating more competitor errors than unrelated errors, but the effect of Segment did not approach significance (coeff. = .02, $z = .27$, $p = .79$), suggesting that vowel and consonant competitor errors were roughly equivalent. This implies that both encoding and recognition in noise may result in a pattern where vowels and consonants are equally important for recognition, which

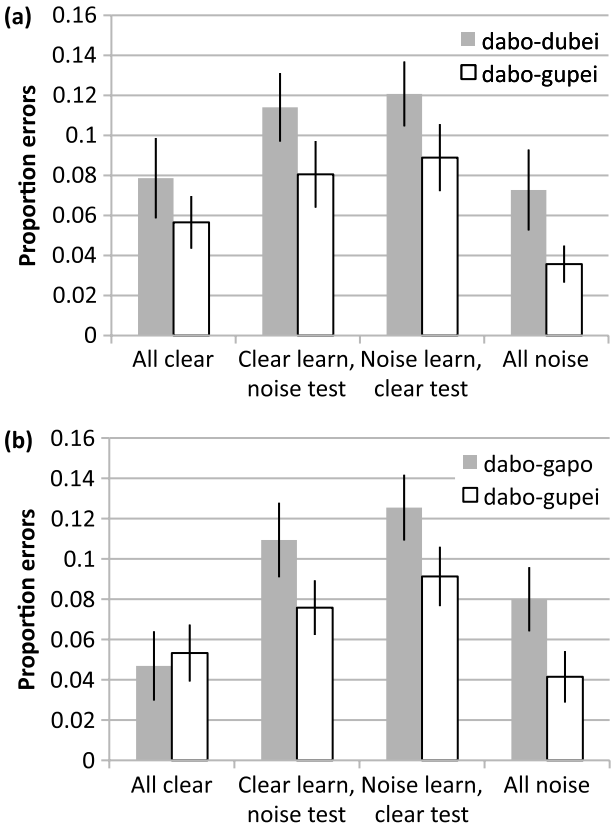

Figure 5. Error data (mean ± standard error) in each condition, for (a) trials with same-consonant competitors, and (b) trials with same-vowel competitors. Unrelated-word errors (white) were divided by two to correct for the greater frequency of unrelated alternatives.

differs from previous findings of a consonant bias in word recognition (Bonatti et al., 2005; Creel et al., 2006; Cutler et al., 2000; Van Ooijen, 1996). This suggestive but inconclusive pattern of results bears exploration in future research.

DISCUSSION

We asked whether listeners' representations of words learned in adverse conditions were specific, and whether good performance in adverse conditions reflected new cue weightings of consonant and vowel information in word recognition. To this end, we had listeners learn new words under clear or noisy conditions, and then tested them under clear or noisy conditions. Listeners benefited from a match in listening conditions between exposure and test: they were both more accurate and faster when exposed and tested under the same noise conditions (both clear or both noisy), suggesting that they formed specific representations of the noisy word forms. There was limited evidence of better performance for learning in the clear (response times only), suggesting that, for the current task, "canonical" listening conditions were not strongly beneficial. Listeners only weakly showed a pattern of increased weighting of vowel information in noisy conditions, though vowels were used just as strongly as consonants for recognition in all conditions but the all-clear condition. Overall, the results of the present study support the hypothesis that listeners readily form acoustically-specific lexical representations in poor listening conditions, but they do not unequivocally support the hypothesis that listeners base these representations on re-weighted consonant and vowel information.

Implications

This research has implications for how listeners not only recognise speech in adverse conditions, but how they represent speech learned under adverse conditions. First, it suggests that listeners do not necessarily encode unfamiliar stimuli as an adjustment to familiar material, but rather as highly-specific, integral representations. It is possible that these representations change over the longer term, perhaps after memory consolidation processes have taken place (e.g. Dumay & Gaskell, 2007) that would allow integration of altered representations with unaltered counterparts. At a more practical level, this implies that listeners attempting to learn words in a novel listening situation—for instance, learning words in a foreign phonological system or a transition from residual low-frequency hearing to a cochlear implant—may perform better by being immersed directly in the new context rather than relying on existing representations. Of course, these real-life situations are more complicated than the current manipulation, which simply requires the listener to recognise preexisting sound categories plus noise, rather than inducing categories not attested in their native language. At the least, though, our data imply that there is more immediate benefit to learning an altered set of stimuli to begin with, rather than learning materials under unaltered conditions and then transitioning to altered conditions.

One question that remains unanswered is whether these specificity effects are stronger for novel words than for familiar words. That is, listeners are better at recognising noisy words when the words were learned in noise. Would they, however, be better at recognising familiar words under adverse conditions than recognising words originally learned in noise? The current data cannot explain whether preexisting "canonical" (relatively noiseless) representations would be more robust to adverse listening conditions than representations learned under adverse listening conditions.

Another unanswered question is the extent to which listeners were utilising representations of preexisting categories. Did participants who learned and were tested in noise show good learning because they were able to noise-shift their speech-sound representations, as in previous demonstrations of phoneme boundary shifts? (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). This would account for the persistence of same-consonant confusions in noisy conditions: listeners were using the consonants, but had shifted their representations of those consonants. This might be thought of as an *extension* of listeners' existing speech sound representations—listeners do not form completely new representations, but they add new variants to existing representations of speech sound patterns. Some evidence exists that listeners extend preexisting phonological knowledge to the recognition of new words: Shatzman and McQueen (2006) found that listeners used their knowledge of Dutch prosody to distinguish newly learned words which were learned without prosodic cues—that is, listeners inferred, in the absence of evidence, that these new words had a familiar prosodic property from generalised prosodic knowledge. This suggests that listeners benefit from being exposed to both familiar and novel contexts (i.e. interdigitated contextual learning).

The results presented in Mattys (2004), and more weakly the current results, suggest an important role for dynamically alterable cue weightings according to listening conditions. On the basis of these results, Mattys argues that we need "an approach that goes beyond assigning absolute weights to individual . . . cues" (p. 405). In the current case, it seems that there is unlikely to be a single consonant–vowel (or auditory-attentional) weighting scheme that is optimal for all listening conditions. Instead, listeners would benefit the most from having flexible cue weighting in word recognition—perhaps acquired by perceptual learning under various listening conditions. Thus, the consonant bias observed in earlier work may simply be one of numerous cue weightings that listeners can implement in good listening conditions. Similar sorts of cue weightings, somewhat analogous to "presets"—existing amplitude profiles in a music player that are optimised for different musical styles—might be used for particular accents, noise types, or vocal idiosyncrasies, allowing the listener to tune to a particular listening context. The process by which learners acquire these "presets" is an intriguing goal for future research.

Finally, our results are overall inconsistent with the hypothesis that consonants are more important for lexical identity than vowels (Bonatti et al., 2005; Nespor et al., 2003). However, it is important to note that our word-learning paradigm differs considerably from learning under natural circumstances. Words are typically embedded in fluent speech rather than being presented in isolation. Thus, we cannot dismiss the more important role for consonants in the task of on-line word recognition. Nevertheless, we did not find conclusive evidence that vowels and consonants are re-weighted during noisy learning. Rather, we found a strong effect of vowel similarity that was indistinguishable from the effect of consonant similarity. To be more specific, our listeners used vowels as strongly as consonants in noisy conditions. Of course, it is entirely possible that this resulted from some unusual property of our stimulus set—we did not find a greater effect of consonant similarity than vowel similarity in clear listening conditions. If there is a consonant bias in clear listening situations, it is likely a learned bias resulting from experience with the greater acoustic differentiability (Macmillan, Goldberg, & Braida, 1988) and greater informational content (Altmann & Carter, 1989) associated with consonants. Of course, calling this a bias assumes that clear listening conditions are the norm, rather than the exception (or one of a range of weighting possibilities). In modern

industrial societies, listeners are exposed to a variety of periodic and aperiodic noise sources, each of which may obscure or distort speech-relevant information in different ways and to different degrees (see Mattys et al., 2009, 2011 this issue). Nonetheless, it would benefit listeners to adapt to specific adverse listening conditions, just as they seem to benefit from adaptation to a variety of accents (e.g. Bradlow & Bent, 2008).

CONCLUSION

In the current study, we explored the specificity and flexibility of listeners' representations of novel words learned under clear and noisy conditions. Listeners were most accurate when tested in the same listening conditions in which they learned the words, suggesting that lexical representations are highly specific to encoding context even when that encoding context obscures the to-be-encoded material. Nonetheless, different listening conditions resulted in equally good recognition of newly learned words. Our data are inconsistent with a hypothesised bias toward consonants over vowels in word recognition, with vowels just as important as consonants in at least a particular type of adverse listening conditions. Whether our listeners' strong use of vowels results from adaptation to listening conditions, from idiosyncrasies of our stimulus set, or from use of isolated words, remains an interesting topic for future exploration. It seems likely that both acoustically-specific learning and flexibility in cue weighting may aid listeners in adapting to changes in listening conditions, including various types of noise or distortion, different talkers, and different accents.

REFERENCES

Altmann, G., & Carter, D. (1989). Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? *Computer Speech and Language*, *3*, 265–275.

Baayen, R. H. (2010). languageR: *Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. R package version 1.0. Retrieved from http://CRAN.R-project.org/package = languageR

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. *Psychological Science*, *16*(6), 451–459.

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

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments & Computers*, *25*, 257–271.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory & Language*, *54*, 1–19.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*, 633–664.

Creel, S. C., & Dahan, D. (2010). The effect of the temporal structure of spoken words on paired-associate learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 110–122.

Creel, S. C., Tanenhaus, M. K., & Aslin, R. N. (2006). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 15–32.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin, Germany: de Gruyter.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, *28*(5), 746–755.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, *116*(6), 3668–3678.

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

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

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time (L). *Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Harrell, F. E. (2009). *Design: Design Package. R package version 2.3-0*. Retrieved from http://CRAN.R-project.org/package = Design

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

Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *The Journal of the Acoustical Society of America*, *49*(6), 1849–1856.

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken word recognition. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1921–1924). Dudweiler: Pirrot.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, *84*(4), 1262–1280.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 397–408.

Mattys, S. L., Bradlow, A. R., Davis, M. H., & Scott, S. K. (2011). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243.

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.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, *54*(2), 205–210.

Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition*, *98*(1), 13–30.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio*, *2*, 221–247.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Pufahl, A., & Samuel, A. G. (2010). *How lexical is the lexicon? Evidence for integrated memory representations.* Poster presented at the 51st annual meeting of the Psychonomic Society, St. Louis, MO.

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

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *106*(6), 3629–3636.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*, 60.
Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.
Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372–377.
Singer, B., & D'Oliveiro, X. (2001). *SpeechSaver*. Retrieved from: http://www.prince-ton.edu/ ~ bdsinger/old/SpeechSaver/
Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4), 487–501.
Van Ooijen, B. (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition*, *24*(5), 573–583.

APPENDIX 1. FOUR ASSIGNMENTS OF PHONEMES TO WORDS

Set A	*Set B*	*Set C*	*Set D*
dabo	puda	gupo	beda
dube	pedo	gepa	budo
dopu	pabe	gobe	bapu
depa	pobu	gabu	bope
bade	dupo	puga	debo
budo	depa	pego	duba
boga	dagu	podu	dage
begu	doge	pade	dogu
gobu	gade	dope	gadu
geba	godu	dapu	gode
gapo	guba	dubo	gepa
gupe	gebo	deba	gupo
poda	bapu	bogu	pabe
pedu	bope	bage	pobu
page	bugo	buda	pego
pugo	bega	bedo	puga

APPENDIX 2. TEST TRIALS (LIST A WORDS)

Segment match	*Target*	*Competitor*	*Distractors*	
Consonants	dabo	dube	poda	pedu
	dube	dabo	pedu	poda
	dopu	depa	pugo	page
	depa	dopu	page	pugo
	bade	budo	geba	gobu
	budo	bade	gobu	geba
	boga	begu	gapo	gupe
	begu	boga	gupe	gapo
	gobu	geba	budo	bade
	geba	gobu	bade	budo
	gapo	gupe	boga	begu
	gupe	gapo	begu	boga
	poda	pedu	dabo	dube
	pedu	poda	dube	dabo
	page	pugo	depa	dopu
	pugo	page	dopu	depa
Vowels	dabo	gapo	poda	boga
	dube	gupe	begu	pedu
	dopu	gobu	page	bade
	depa	geba	budo	pugo
	bade	page	gobu	dopu
	budo	pugo	geba	depa
	boga	poda	gapo	dabo
	begu	pedu	gupe	dube
	gobu	dopu	page	bade
	geba	depa	budo	pugo
	gapo	dabo	poda	boga
	gupe	dube	begu	pedu
	poda	boga	gapo	dabo
	pedu	begu	dube	gupe
	page	bade	gobu	dopu
	pugo	budo	depa	geba