Perception & Psychophysics
2008, 70 (4), 604-618
doi: 10.3758/PP.70.4.604

Does perceptual learning in speech reflect changes in phonetic category representation or decision bias?

Constance M. Clarke-Davidson, Paul A. Luce, and James R. Sawusch
University at Buffalo, State University of New York, Buffalo, New York

Recent studies show that perceptual boundaries between phonetic categories are changeable with training (Norris, McQueen, & Cutler, 2003). For example, Kraljic and Samuel (2005) exposed listeners in a lexical decision task to ambiguous /s–ʃ/ sounds in either s-word contexts (e.g., *legacy*) or ʃ-word contexts (e.g., *parachute*). In a subsequent /s/–/ʃ/ categorization test, listeners in the /s/ condition categorized more tokens as /s/ than did those in the /ʃ/ condition. The effect—termed *perceptual learning in speech*—is assumed to reflect a change in phonetic category representation. However, the result could be due to a decision bias resulting from the training task. In Experiment 1, we replicated the basic Kraljic and Samuel (2005) experiment and added an AXB discrimination test. In Experiment 2, we used a task that is less likely to induce a decision bias. Results of both experiments and signal detection analyses point to a true change in phonetic representation.

Recent research on speech perception has focused on perceptual learning as a means of solving the problem of variability in speech. According to the perceptual learning account, the idiosyncrasies of a speaker's production of speech sounds are learned and retained rather than mapped onto more abstract representations and discarded. An example of a talker-specific characteristic that might be handled by perceptual learning comes from Newman, Clouse, and Burnham (2001). They showed that the acoustics of /s/ (as in *see*) and /ʃ/ (as in *she*) vary greatly from speaker to speaker, and that, across speakers, there is extensive overlap between /s/ and /ʃ/ in the mean frequency of frication noise. However, within a given speaker, the /s/–/ʃ/ overlap is greatly reduced. With some experience, a listener could learn a speaker's characteristic /s/ and /ʃ/ frequency ranges and distinguish them in a speaker-specific manner.

A variety of studies have demonstrated the beneficial effect of experience with talker-specific characteristics. Nygaard and colleagues (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994) showed improved word identification in noise for voices on which listeners had been trained for several days, and McGarr (1983) found that experienced listeners transcribed the speech of deaf speakers more accurately than did inexperienced listeners. Even very brief exposure to a speaker has been shown to affect perception. In a classic study, Ladefoged and Broadbent provided an early demonstration of rapid adaptation to speaker characteristics (Ladefoged, 1989; Ladefoged & Broadbent, 1957) in which the formant ranges of a precursor sentence determined the vowel perceived in a /b/–V–/t/ test word. More recently, Clarke and Garrett (2004) found an increase in processing efficiency for Spanish- and Chinese-accented speech after less than a minute of exposure.

There is evidence that the benefit of previous experience with a talker is at least in part due to flexibility at the phonetic level of processing. Norris, McQueen, and Cutler (2003) found a shift in Dutch listeners' /s/–/f/ categorization boundary depending on whether an ambiguous /s/–/f/ sound replaced /s/s or /f/s in a preceding lexical decision task. If the ambiguous sound occurred in s-final word contexts, then more /s/ responses were given in the categorization test, and vice versa for the f-final word condition. Importantly, there were only 20 samples of the ambiguous sound in the lexical decision task, again demonstrating that perceptual learning in speech can occur quite rapidly. The authors concluded that feedback from lexical representations altered the /s/ and /f/ phonetic categories. Subsequent studies have added to our understanding of perceptual learning in speech. The learning is speaker specific in cases of spectral contrast, such as /s/ versus /f/ and /s/ versus /ʃ/ (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007), but generalizes across speakers for nonspectral contrasts, as in the timing contrast in /t/ versus /d/ (Kraljic & Samuel, 2006, 2007). McQueen, Cutler, and Norris (2006) found the perceptual learning effect to generalize to words not heard in training, suggesting a sublexical locus of the effect. Kraljic and Samuel (2005) replicated the effect in English with the /s/–/ʃ/ contrast and found it to be robust to both time delays and corrective input from the same speaker. Even when testing is delayed by 12 h, the effect remains robust (Eisner & McQueen, 2006). Finally, lexical knowledge is

C. M. Clarke-Davidson, cmclarke@ualberta.ca

not the only type of disambiguating information that can drive perceptual learning. Comparable phonetic boundary shifts have been shown using visual speech as the disambiguating information (e.g., an ambiguous sound between /t/ and /p/ synchronized with a face producing an unambiguous /p/; Bertelson, Vroomen, & de Gelder, 2003; van Linden & Vroomen, 2007).

Shifts in categorization boundaries as a function of perceptual learning have been taken as evidence for changes in phonetic representations. However, an alternative explanation exists: The effect could result from a decision bias developed during the training task. In a majority of published studies of the effect, the training task was auditory lexical decision. In lexical decision, listeners must implicitly assign the ambiguous sound to one category or the other in order to make the word–nonword decision. Instead of causing a remapping of the acoustic–phonetic input to the phonetic categories, as has been assumed, the lexical feedback could result in an implicit decision criterion shift to use, for example, more /s/ labels, without any underlying remapping. Such a bias would help participants make faster word decisions in the ambiguous cases. We define bias as the increased likelihood to give a particular response—such as /s/—given any acoustic input, or the need for less evidence for a particular response. Bias contrasts with remapping, in which a region of acoustic–phonetic space formerly associated with one category is now associated with another. In either the remapping or the bias case, a categorization boundary shift would be predicted. Given that the perceptual learning effect has had a substantial impact on the field of spoken word recognition (supporting the view of speech perception as flexible and dynamic), we believe that it is important to further explore this effect to be more confident in the inferences being made regarding the underlying mechanisms responsible for learning.

In this study, we used two experimental methods as well as signal detection analysis to examine the possibility that a decision bias is involved in the perceptual learning effect. In Experiment 1, we tested the effect using an AXB discrimination task in addition to the usual categorization task. Perceptual discrimination is typically better for two tokens that fall on opposite sides of a category boundary—as compared with tokens that fall within a phonetic category—resulting in a discrimination peak near the boundary (Liberman, Harris, Hoffman, & Griffith, 1957). This occurs because of the use of phonetic categories in making discrimination judgments (assuming there are no auditory discontinuities along the acoustic continuum). If the underlying phonetic category boundaries change, then we expect the location of peak discrimination to change accordingly. Therefore, a lack of shift in the discrimination peak would suggest that there is no underlying change in the phonetic category representations. Although performance is based on phonetic categories for both categorization and discrimination tasks, their results could differ if the categorization effect is due entirely to decision bias and the discrimination task does not involve this bias.

In Experiment 2, we tested the bias hypothesis by using a task other than lexical decision to expose listeners to the ambiguous sounds. Instead of lexical decision, we used a same–different discrimination task in which listeners heard two items (words or nonwords) and decided whether they were the same or different. In this task, no decision about the ambiguous sound was required because the items in each pair were either identical or completely different phoneme sequences. This task should reduce the likelihood of a decision bias because no labeling decision is required. We note, however, that since the collection of these data, three articles have been published reporting replication of the perceptual learning effect using training tasks other than lexical decision. In Eisner and McQueen (2006), listeners were exposed to ambiguous tokens while passively listening to a story. McQueen, Norris, and Cutler (2006) used a trial counting task during training, and Leach and Samuel (2007) used an old–new recognition task. Presumably, these training tasks also reduce the likelihood of developing a decision bias. However, the results of our study are valuable for two reasons. First, although one can speculate on the motivational and processing dynamics that a certain experimental task induces, these speculations can be incorrect. Our study used experimental and statistical techniques to tease apart representational and bias changes in the perceptual learning effect. In addition to adding the discrimination task, we conducted signal detection analyses on the categorization data to explicitly test for sensitivity and decision criterion differences. In fact, the signal detection analysis revealed that the use of a task that does not require an explicit word–nonword decision does not preclude the development of a decision bias. The second contribution of our study is the replication of the perceptual learning effect within a new training context and with a new dependent measure, which strengthens confidence in the generalizability of the effect.

It might also be argued that McQueen, Cutler, and Norris's (2006) replication of the perceptual learning effect using a priming task at test rules out bias as a factor because explicit phoneme categorization was not involved and because the effect generalized to new words. Participants were trained with ambiguous /s/–/f/ sounds in an auditory lexical decision task, and the test phase was visual lexical decision with auditory primes. Ambiguous primes (e.g., /nai?/) facilitated responses for the /f/ version of the target (e.g., *knife*) for f-trained participants and the /s/ version (e.g., *nice*) for s-trained participants. However, if a bias had developed in the training task, it could be active within the phonetic processing system itself and therefore affect the processing of subsequent words, regardless of the response required and regardless of whether the words were heard in training. In general, no single experiment can unequivocally determine the source of the perceptual learning effect. Rather, convergent evidence is required, and the present study offers an explicit test of whether the effect is based on representational changes or bias.

In both experiments, we used the materials and basic design of Kraljic and Samuel (2005). In order to replicate the basic perceptual learning effect, we duplicated Kraljic and Samuel's (2005) Experiment 1, Phases I (lexical decision) and III (/s/–/ʃ/ category identification) with a male voice. This simplified version of Kraljic and Samuel's (2005) experiment allowed us to replicate the categoriza-

tion boundary shift, a prerequisite to testing for a corresponding discrimination peak shift.

EXPERIMENT 1

The purpose of this experiment was to replicate the perceptual learning effect with the /s/–/ʃ/ contrast and to test for an accompanying change in discrimination ability. Listeners were exposed to sounds that were ambiguous between /s/ and /ʃ/ in a lexical decision task. In the s-training condition, an ambiguous sound replaced the /s/ in 20 words (e.g., /lɛgəʔi/ *legacy*); in the ʃ-training condition, an ambiguous sound replaced the /ʃ/ in 20 words (e.g., /pɛrəʔut/ *parachute*). Two tests followed the lexical decision task: a phonetic categorization test, in which listeners categorized tokens from an /ɑsi/–/ɑʃi/ continuum, and an AXB discrimination test, in which they discriminated token pairs from the same /ɑsi/–/ɑʃi/ continuum. If training with the ambiguous sounds alters the /s/ and /ʃ/ category representations so that the acoustic–phonetic space is remapped, then the peak in discrimination should shift along with the categorization boundary.

Method

Participants

In both Experiments 1 and 2, listeners (1) had parents whose native language was American English, (2) had no exposure to any other language before age five, (3) were not fluent in any other language, (4) reported no current speech or hearing disorders, and (5) were right-handed. A total of 121 University at Buffalo undergraduates participated in Experiment 1 for course credit. Of those, 33 were excluded for the following reasons: 19 did not meet the participation criteria, 12 had high error rates or no-response rates (see the Results section), and 2 did not finish in the time allowed. The final sample consisted of 88 listeners (47 males, 41 females), half in each of the s- and ʃ-training conditions.

Materials

The materials for the lexical decision task were taken from Appendix A of Kraljic and Samuel (2005) and consisted of 20 critical s-words, 20 critical ʃ-words, 60 filler words, and 100 filler nonwords. The critical s- and ʃ-words had between two and five syllables and contained a syllable-initial /s/ or /ʃ/ late in the word. The s-words had a mean of 3.2 syllables, a mean word frequency of 17.9/million (Kučera & Francis, 1967), and did not contain /ʃ/, /z/, /ʒ/, or any other /s/. The ʃ-words had a mean of 3.2 syllables, a mean word frequency of 22.8/million, and did not contain /s/, /z/, /ʒ/, or any other /ʃ/. The 60 filler words did not contain /s/, /ʃ/, /z/, or /ʒ/ and were similar to the critical words in mean length (3.1 syllables) and mean frequency (13.2/million). The 100 filler nonwords were based on the 60 filler words, and 40 other words not used in the study. They were created by Kraljic and Samuel (2005) by changing several phonemes in each word to other phonemes with the same manner of articulation. (One repeated nonword, *galliwinou*, in Kraljic and Samuel [2005] was replaced with the nonword *aginode*.) The nonwords also did not contain /s/, /ʃ/, /z/, or /ʒ/; however, we discovered one containing /s/ (*bawaseet*) and removed it from the materials midway through the experiment.

Stimulus Construction

Recording. The words, nonwords, and /ɑsi/ and /ɑʃi/ tokens were produced by a linguistically trained male native speaker of American English. For the nonwords, we created an IPA transcription based on the standard orthography given in Kraljic and Samuel (2005) for use by the speaker during recording. The speech was recorded on a CD (44.1 kHz, 16 bit) in a quiet room. The stimulus items were copied to individual files; silence at the beginning and end was trimmed, and the files were peak normalized to 90% of maximum amplitude resolution.

Lexical decision stimuli (for exposure task). For each critical word, the speaker recorded an /s/ version and an /ʃ/ version (e.g., *legacy* and *legashy*, *parachute* and *parasute*) so that a unique ambiguous fricative could be created for each one (following Kraljic & Samuel, 2005). The /s/ and /ʃ/ portions of the waveforms were copied into their own files, equated on number of samples, and blended together at five ratios. The first and second authors independently listened to the isolated fricative blends for each critical word and chose the most ambiguous. If they agreed, then that blend was used; if they chose ratios that were one step apart, then an intermediate blend was created; if they chose ratios that were two steps apart, the blend in between was used. The chosen fricative blend was set to the mean dB of the original /s/ and /ʃ/, and the beginning and end of the /s/ version of the word were added to it (e.g., $lega_{s}$–?–y_{s}) to create the whole word.

/asi/–/aʃi/ continuum (for categorization and discrimination tasks). The /ɑsi/–/ɑʃi/ continuum was created by blending the /s/ and /ʃ/ portions of natural /ɑsi/ and /ɑʃi/ recordings. The fricative portions of the waveforms were copied to separate files, equated on number of samples, blended together at ratios between 5% s/95% ʃ and 95% s/5% ʃ in 5% steps, and set to the mean dB of the original /s/ and /ʃ/ tokens. The /ɑ/ and /i/ from the original /ɑsi/ token were then added to all blends. All waveform manipulations were performed using Praat (Boersma & Weenink, 2006) and Peak (version 4.13, BIAS, Inc.) waveform editing programs. On the basis of pilot categorization tests, seven tokens were chosen to range from a good /s/ to a good /ʃ/ (% s/% ʃ in blend): 35/65, 45/55, 55/45, 65/35, 75/25, 85/15, 95/5. A greater ratio of /s/ to /ʃ/ in the blend was required for listeners to give a high proportion of /s/ responses because the /ʃ/ frication tended to dominate perceptually.

Procedure

Between 1 and 4 participants were tested at a time in a quiet room. Each sat in front of a computer screen and a button box and heard all of the stimuli at a comfortable listening level over headphones. Stimulus presentation and response recording were controlled by PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). The lexical decision task was presented first, followed by the categorization and discrimination tasks, which were counterbalanced to control for any possible carryover task effects. In the s-training group, 24 listeners performed the categorization task before the discrimination task, and 20 performed them in the reverse order. In the ʃ-training group, 20 performed categorization before discrimination, and 24 performed the reverse. Between tasks, the experimenter placed the appropriate button labels for the upcoming task on the button box and avoided speaking in order not to give any interfering speech input. For all tasks, each trial began 2,000 msec after the previous buttonpress; if no response was given within 4,000 msec, the next trial was initiated.

For the lexical decision task, listeners were instructed to press the button labeled "Word" (right hand) if they heard a real English word and the button labeled "No word" (left hand) if they heard a nonsense word. Both speed and accuracy were emphasized. The 100 words and 100 nonwords were presented in a different random order for each participant.

Although listeners were given general instructions about the categorization task at the beginning of the experiment, they were not told what sounds they would hear so that they would not be alerted to the manipulation in the lexical decision task. Detailed instructions were given on-screen immediately before the task began. The seven tokens from the /ɑsi/–/ɑʃi/ continuum were presented in 10 randomized blocks following 1 block of practice. Listeners were instructed to press the button labeled "AH–SEE" or the button labeled "AH–SHE" to indicate what they heard. Button labels were counterbalanced across participants. Both speed and accuracy were emphasized.

For the AXB discrimination task, detailed instructions were also only given immediately before the task began. On each trial, three

Table 1
Lexical Decision (Experiment 1) and Same–Different (Experiment 2) Accuracy and Reaction Times (RTs, in Milliseconds) for Critical s- and ʃ-Words

	s Training				ʃ Training			
	/?s/		/ʃ/		/s/		/?ʃ/	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				Experiment 1				
Percent correct	98.64	3.12	99.55	1.45	99.43	1.61	92.84	7.66
RT	250	137	206	143	235	120	287	130
				Experiment 2				
Percent correct	98.86	2.64	99.09	2.51	99.32	1.76	99.09	2.51
RT	745	213	820	250	835	261	806	273

Note—/?s/ and /?ʃ/, respectively, denote /s/ and /ʃ/ words containing ambiguous sounds. Percent correct refers to percent *word* response for Experiment 1 and percent *different* response for Experiment 2. RTs are for correct trials and were measured from the stimulus offset in Experiment 1 and from the onset of the second stimulus of a pair in Experiment 2.

tokens from the /ɑsi/–/ɑʃi/ continuum were presented sequentially. The first and third tokens were always different (and constituted a pair), and the second was the same as the first or third. Listeners were instructed to indicate which of the flanking stimuli matched the middle token by pressing a button on the button box. A mix of one-step and two-step pairs was presented. There were 6 one-step pairs consisting of tokens lying next to each other on the continuum, and 5 two-step pairs consisting of tokens lying two steps apart on the continuum. Both types were included because the one-step pairs provided a finer grained measure of discrimination ability; however, in pilot testing, participants found them very difficult. The two-step pairs were included in the case that performance on the one-step pairs was at chance, as well as to provide easier trials for motivational purposes. Each of the 11 pairs was presented in four triads (AAB, ABB, BBA, BAA), and there were four randomized blocks of the 44 triads, resulting in 176 trials total. A break was provided after the second block. A varied selection of 5 two-step triads was presented as practice. Participants were instructed that some trials would be very difficult, but that they should respond as accurately as possible and guess if necessary. Each trial began with "READY" displayed on the screen for 1,000 msec, followed 500 msec later by the three tokens in sequence, separated by 500 msec. Listeners pressed either the button labeled "2nd is identical to the 1st" (left hand) or the button labeled "2nd is identical to the 3rd" (right hand).

Results and Discussion

We first evaluated lexical decision performance to see whether the critical items were labeled as words. Two listeners were excluded because their accuracy for the filler nonwords was less than 75%. Reaction times (RTs) for correct (*word*) responses to the critical items were calculated from word offset because stimulus durations were not equated across ambiguous and natural conditions. Table 1 shows percent correct and RT for the critical words, collapsed across test order. Accuracy was high for both training conditions, indicating that despite the presence of the ambiguous /s/–/ʃ/ sounds, the critical items were generally heard as words. We ran an ANOVA on percent correct with critical word (s-words vs. ʃ-words) as a within-participants factor and training condition (s vs. ʃ) and test order (categorization first vs. discrimination first) as between-participants factors. There were significant main effects of critical word [$F(1,84) = 19.64$, $p < .001$, $\eta^2_G = .10$ (generalized eta-squared; Bakeman, 2005; Olejnik & Algina, 2003)] and training condition [$F(1,84) = 21.39$, $p < .001$, $\eta^2_G = .11$]; however, these were mediated by a significant interaction between critical word and training condition [$F(1,84) = 33.78$, $p < .001$, $\eta^2_G = .17$], reflecting the ʃ-trained listeners' lower accuracy for the ambiguous ʃ-words. The ʃ-trained listeners were also slower to respond to the ambiguous ʃ-words, as indicated by a significant interaction between critical word and training condition in the RT analysis [$F(1,84) = 52.90$, $p < .001$, $\eta^2_G = .03$]. Post hoc comparisons showed that responses to ambiguous ʃ-words were slower and less accurate than to natural ʃ-words [RT, $t(86) = 2.78$, $p < .01$, Cohen's $d = 0.60$; percent correct, $t(46.1) = 5.70$, $p < .001$, Cohen's $d = 1.23$ (equal variances not assumed)]. There were no significant differences in accuracy or RT between ambiguous and natural s-words.

The lexical decision results indicate that the ambiguous ʃ-words were not as acceptable as words as were the ambiguous s-words. However, the overall acceptance rate for the ambiguous ʃ-words was quite high (92.84%), indicating that the majority of critical trials provided the opportunity for learning that the ambiguous sound belonged to the /ʃ/ category.

Following the experiment, some listeners were asked whether they noticed anything unusual about the words in the lexical decision task. Of the 60 asked, 2 (1 in s training, 1 in ʃ training) reported hearing the manipulation. When asked further whether they noticed anything about the /s/ or /ʃ/ sounds in particular, an additional 10 (4 in s training, 6 in ʃ training) gave a variety of responses, such as "softer," "pronounced differently," "blended together," "sounded the same," and "slurred together." (We should note that although the experimenter asked specifically about the lexical decision task, we suspect that some responses were made on the basis of the categorization or discrimination tests.) There were no changes in the patterns of categorization or discrimination results (see below) when these participants were excluded from the sample.

Results for the /ɑsi/–/ɑʃi/ categorization task are shown in Figure 1. The total percent /ɑʃi/ responses was calculated for each listener. Trials with RTs less than 200 msec or greater than 2,000 msec were excluded from analysis. There was a sizeable difference in the /s/–/ʃ/ boundary

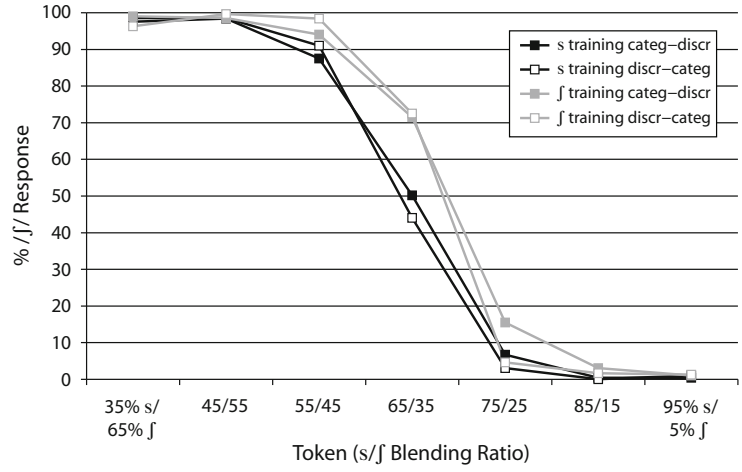

Figure 1. Experiment 1 categorization results displayed by training condition and categorization–discrimination test order. Categ, categorization; discr, discrimination.

for the two training conditions in the expected direction. Listeners who had been trained that the ambiguous sounds belonged to the /ʃ/ category gave more /ɑʃi/ responses (M = 53.96%, SD = 5.38) than those who had been trained that they belonged to the /s/ category (M = 48.42%, SD = 5.40; M_{Diff} = 5.54, 95% CI_{Diff} = [3.24, 7.84]). This was confirmed in a 2 (training condition) × 2 (categorization–discrimination test order) ANOVA on total percent /ɑʃi/ response, which revealed a significant main effect of training condition [$F(1,84) = 23.45$, $p < .001$, $\eta^2_G = .22$]. There was no main effect of or interaction with test order, suggesting that the perceptual learning effect was the same whether the categorization test occurred immediately after exposure or after the intervening discrimination test. The robustness of the perceptual learning effect to intervening speech input is consistent with previous findings (Eisner & McQueen, 2006; Kraljic & Samuel, 2005; however, see Kraljic & Samuel, 2006; van Linden & Vroomen, 2007).

Replication of the perceptual learning effect in the categorization task allows us to move to the central question of this experiment: Does discrimination ability change in accord with the shift in the categorization boundary? The AXB discrimination task tested this question. For each listener, the percent-correct response was calculated for each discrimination pair across the 16 presentations (4 triads × 4 blocks) to which there was a response. Responses were not counted if they occurred before the beginning of the third token of a triad. Ten participants were excluded from the experiment for responding before this point or not responding at all on more than 10% of discrimination trials. This somewhat high exclusion rate is presumably due to the difficulty of the discrimination task.

As shown in Figure 2A for the one-step pairs, performance at the endpoints of the continuum is at chance level (50%), whereas accuracy nears 70% in the middle. This pattern is typical for a phonetic discrimination task and presumably reflects the involvement of phonetic category representations. However, there is a clear difference in the accuracy peak for the two training conditions. The ʃ-trained groups were best able to discriminate the 65/35 and 75/25 tokens, whereas the s-trained groups' highest accuracy was for the 55/45 and 65/35 tokens. A similar pattern can be seen in Figure 2B for the two-step pairs. Overall performance is better because the paired tokens were further apart on the continuum, but accuracy is still worse at the continuum ends than in the middle. Although all groups were best able to discriminate the 55/45 and 75/25 tokens, the shape of the curves is consistent with their categorization performance. Specifically, for the s-training conditions, the second best discriminated pair was 45/55 and 65/35, but for the ʃ-training conditions, the second-best discriminated pair was 65/35 and 85/15.

In order to test these patterns statistically, we conducted separate ANOVAs for the one-step and two-step sets, each with training condition and test order as between-participants factors, and token pair as a within-participants factor. For both sets, there was a main effect of token pair [$F_{1\text{-step}}(5,420) = 23.65$, $p < .001$, $\eta^2_G = .19$; $F_{2\text{-step}}(3.82, 320.73) = 118.86$, $p < .001$, $\eta^2_G = .48$; Huynh–Feldt corrected for nonsphericity], reflecting the worse accuracy at the ends of the continuum as compared with the middle. There were also significant interactions between token pair and training condition [$F_{1\text{-step}}(5,420) = 5.12$, $p < .001$, $\eta^2_G = .05$; $F_{2\text{-step}}(3.82, 320.73) = 3.98$, $p < .01$, $\eta^2_G = .03$; Huynh–Feldt corrected for nonsphericity], verifying the difference in discrimination ability along the continuum, depending on training. The only effect involving test order was an interaction with training condition in the two-step set [$F_{2\text{-step}}(1,84) = 4.28$, $p < .05$, $\eta^2_G = .02$]. As can be seen in Figure 2B, the ʃ-trained group that performed the discrimination task immediately after the lexical decision task was more accurate overall than the ʃ-trained group that performed the categorization task first. There was no such difference for the s-trained groups. We have no explanation for this effect at this time; because it

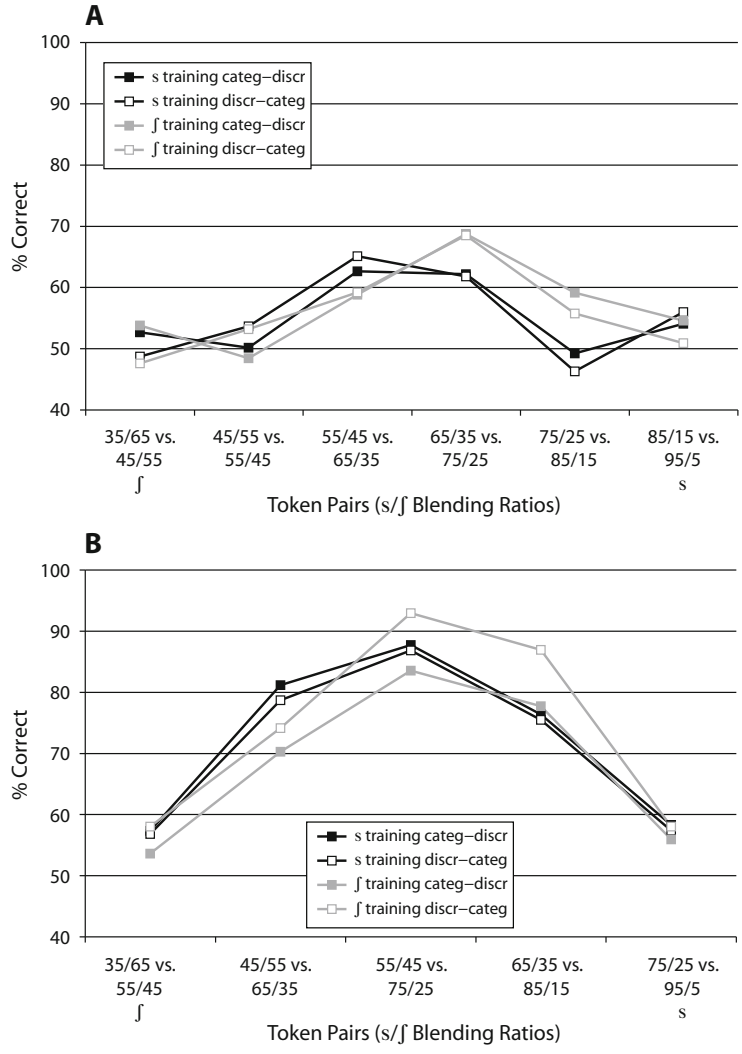

Figure 2. Experiment 1 AXB discrimination accuracy for (A) one-step and (B) two-step trials displayed by training condition and categorization–discrimination test order. One- and two-step trials were mixed within the discrimination test. Chance is 50%. Categ, categorization; discr, discrimination.

did not interact with token pair and does not compromise the results or conclusions, we did not explore it further.

The results of the discrimination test are consistent with a change in phonetic category representation—that is, a remapping of acoustic–phonetic space to the /s/ and /ʃ/ phonetic categories. If the AXB task is performed primarily by assigning each of the fricatives in the triad to a phonetic category and comparing the category of the second with those of the first and third, the discrimination functions seem to follow directly from the categorization results. This observation is confirmed by the predicted discrimination scores shown in Figure 3 for the two training conditions (collapsed across test order). The predicted scores were derived using the following formula: predicted discrimination accuracy $= 0.5 + 0.5[p(T_1) - p(T_2)]^2$, where $p(T_1)$ and $p(T_2)$ are the proportions of /ʃ/ categorization response, respectively, for tokens T_1 and T_2 of a pair. These predictions assume that the listeners used a phonetic strategy (i.e., perception is categorical; Liberman et al., 1957; Pollack & Pisoni, 1971). Actual performance is generally better than predicted, but the overall patterns are the same.

The purpose of Experiment 1 was to test whether discrimination performance would change along with categorization, as would be predicted if the underlying category representations were modified with training. Discrimination performance did indeed change in the predicted way. The results are consistent with the hypothesis that training retunes listeners' /s/ and /ʃ/ category representations, and they support previous interpretations of the perceptual

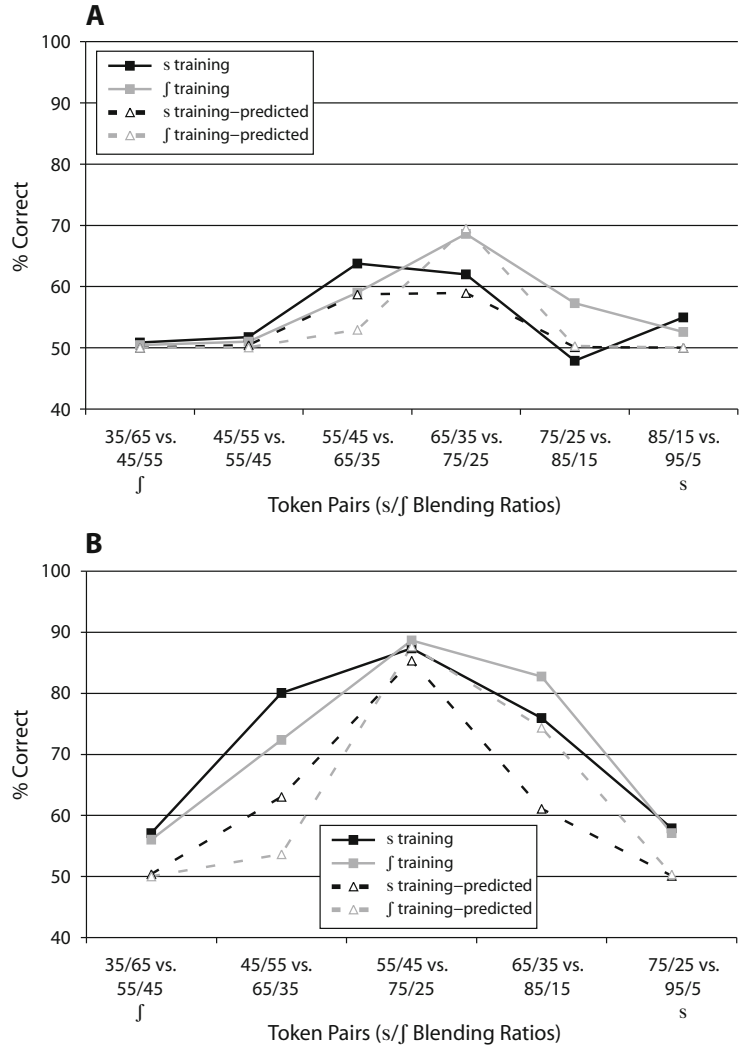

Figure 3. Experiment 1 AXB discrimination accuracy (collapsed across test order) and predicted accuracy on the basis of categorization data for (A) one-step and (B) two-step trials. Chance is 50%.

learning effect (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Norris et al., 2003). The purpose of Experiment 2 was to further test the bias hypothesis by changing the potential source of the bias—the training task. Specifically, we changed the exposure task to one that would be less likely to cause a decision bias: a same–different task. Replicating the categorization boundary shift (and discrimination peak shift) would provide further evidence for a representational change.

EXPERIMENT 2

Instead of using lexical decision to expose listeners to the ambiguous /s/–/ʃ/ stimuli, we used a simple same–different discrimination task that included the same words and nonwords from Experiment 1. On each trial, two items were presented that were either the same (the same word or nonword) or different (different words or nonwords or a word and a nonword). Importantly, the same–different task does not require a decision about the phonetic identity of the ambiguous sound. Developing a bias regarding the /s/ and /ʃ/ categories would not benefit task performance as it would for lexical decision; therefore, it seems less likely to occur. As in Experiment 1, both categorization and discrimination tests followed the exposure task.

Method

Participants

A total of 59 University at Buffalo undergraduates participated in Experiment 2 for course credit. Of those, 15 were excluded for the following reasons: Eight did not meet the participation criteria, 5 had high no-response rates (see the Results section), 1 did not have a well-defined /s/–/ʃ/ boundary, and 1 because all conditions were

filled. The final sample was 44 listeners (25 males, 19 females), half in each of the s- and ʃ-training conditions.

Materials and Stimuli

The same–different exposure test set consisted of 140 pairs: 70 same (35 word–word, 35 nonword–nonword), and 70 different (20 $word_1$–$word_2$, 15 $nonword_1$–$nonword_2$, 20 word–nonword, 15 nonword–word). The word and nonword stimuli were the same as those used in the lexical decision task in Experiment 1. They included 20 critical s-words, 20 critical ʃ-words, 60 filler words (plus 10 repeated), and 99 nonwords (plus 1 repeated). Ten of the critical s-words and 10 of the critical ʃ-words were the first items in the $word_1$–$word_2$ pairs; the remaining 10 of each were the first items in the word–nonword pairs. Therefore, the critical words were heard only once, as in Experiment 1, and were always presented first in a pair in order to ensure that the listeners heard the whole word before responding. The repeated nonword, *galliwinou*, was included, and the nonword *aginode* replaced *bawaseet*. Ten filler words were repeated in order to have a balanced set. The /ɑsi/–/ɑʃi/ continuum used in the categorization and discrimination tests was identical to that in Experiment 1.

Procedure

Because there were no relevant test-order effects in Experiment 1, all participants received the same order of events: same–different exposure task, categorization test, and discrimination test. On each same–different trial, two items were presented over headphones with an interstimulus interval of 500 msec. Participants were instructed to press the button labeled SAME (right hand) if the items were the same or the button labeled DIFFERENT (left hand) if they were different. Both speed and accuracy were emphasized. Item pairs were presented in a different random order for each participant. The duration of the same–different block was approximately the same as the duration of the lexical decision block in Experiment 1. All other aspects of the experiment were the same as those in Experiment 1.

Results and Discussion

Accuracy for the same–different task was high, with an overall mean over 98% for both training conditions. The RT was measured from the beginning of the second item of each pair. Table 1 shows the percent correct and RTs (correct responses only) for the critical words. We ran separate ANOVAs on percent correct and RT, each with critical word (s-words vs. ʃ-words) as a within-participants factor and training condition (s vs. ʃ) as a between-participants factor. There were no statistically significant effects in the accuracy analysis. However, for the RT analysis, there was a significant interaction between critical word and training condition [$F(1,42) = 17.71$, $p < .001$, $\eta^2_G = .02$]. This result appears to reflect the faster RTs for pairs containing an ambiguous /s/ word as compared with a natural /s/ and the lack of difference between the ambiguous and natural /ʃ/ words. However, post hoc tests revealed that the only significant pairwise difference was between the faster (ambiguous) /s/ and slower (natural) /ʃ/ critical words for the s-training condition [$t(21) = 4.34$, $p < .001$, Cohen's $d = 0.32$]. It is not clear why *different* responses would be faster for pairs containing a word with an ambiguous sound as compared with the natural case. In particular, if listeners used the lexical level of processing to make the same–different judgment, then one would expect slower responses in the ambiguous case because lexical access would be slowed. Therefore, lexical involvement in this task is not strongly indicated (see also Vitevitch & Luce, 1999). Regardless of the basis of these results, they indicate that the ambiguous items did not interfere with the task and suggest that there was little need for listeners to develop a labeling bias.

When asked whether they noticed anything unusual about how the words were pronounced in the lexical decision task, 1 of the 59 participants reported hearing the manipulation (in the s-training condition). When further probed about the /s/ and /ʃ/ sounds in particular, 8 additional participants (4 in s training, 4 in ʃ training) responded positively, again with most saying that the sounds "sounded similar" or were "slurred together." (As in Experiment 1, we suspect some responses were regarding the categorization or discrimination tests.) There were no changes in the patterns of categorization or discrimination results (see below) when these participants were excluded from the sample.

The /ɑsi/–/ɑʃi/ categorization data were processed in the same way as they were in Experiment 1. One participant was excluded for not having a well-defined /s/–/ʃ/ boundary. As shown in Figure 4, the ʃ-trained participants categorized more tokens as /ʃ/ ($M = 58.55\%$, $SD = 8.29$) than did the s-trained participants ($M = 47.71\%$, $SD = 7.61$; $M_{Diff} = 10.84$, 95% $CI_{Diff} = [6.00, 15.69]$) [$F(1,42) = 20.40$, $p < .001$, $\eta^2_G = .33$]. Thus, the perceptual learning effect was replicated with the same–different exposure task with an even larger effect size.

The AXB discrimination data were processed in the same way as they were in Experiment 1. Five participants were excluded for responding before the third token of a triad or not responding at all on more than 10% of the trials. Figure 5 shows the actual and predicted results for the one-step and two-step sets. As expected, the main effect of token pair was significant [$F_{1\text{-step}}(5,210) = 13.80$, $p < .001$, $\eta^2_G = .19$; $F_{2\text{-step}}(3.77, 158.57) = 63.03$, $p < .001$, $\eta^2_G = .49$; Huynh–Feldt corrected for nonsphericity]. Also similar to Experiment 1, discrimination performance across the /ɑsi/–/ɑʃi/ continuum varied depending on the training condition. For the one-step pairs (Figure 5A), ʃ-trained participants showed a peak in discrimination for the 65/35–75/25 pair, whereas s-trained participants performed similarly on the 55/45–65/35 and 65/35–75/25 pairs. In line with these observations, token pair interacted significantly with training condition [$F_{1\text{-step}}(5,210) = 2.26$, $p < .05$, $\eta^2_G = .04$]. The pattern of the predicted values is similar to that of the actual values. The two-step results (Figure 5B) are also similar to those of Experiment 1. Participants in both training conditions discriminated the 55/45–75/25 pair most accurately, but the next most accurate pair was closer to the /s/ end of the continuum for the ʃ-trained group, and closer to the /ʃ/ end for the s-trained group. However, the token pair by training-condition interaction did not reach statistical significance for the two-step set [$F_{2\text{-step}}(3.77, 158.57) = 2.02$, $p = .098$, $\eta^2_G = .034$; Huynh–Feldt corrected for nonsphericity]. Interestingly, the predicted discrimination accuracies show a greater difference between the training conditions than do the actual accuracies (Figure 5B). On the basis of the categorization data, the discrimination peak is predicted to be at the 65/35–85/15 pair for the ʃ-trained group, and at

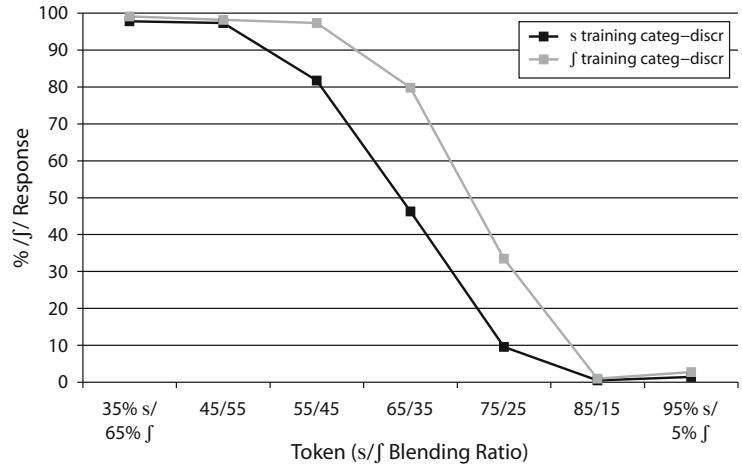

Figure 4. Experiment 2 categorization results displayed by training condition. Categ, categorization; discr, discrimination.

the 55/45–75/25 pair for the s-trained group. Instead, both groups' peaks are at the 55/45–75/25 pair. The discrimination results, therefore, suggest less difference between the groups than the categorization results do.

The overall results of Experiment 2 are somewhat more complicated than those of Experiment 1. The /s/–/ʃ/ categorization boundary shift was replicated using a new exposure task that we expected to be less likely to encourage a decision bias than the lexical decision task. In addition, the discrimination peak shift again accompanied the categorization boundary shift for the one-step trials, replicating Experiment 1. However, the training conditions did not differ significantly for the two-step discrimination trials, even though they were predicted to differ substantially on the basis of the categorization data. Although the weaker effect for the two-step discrimination suggests that the difference between the training conditions may not be as extensive as the categorization results indicate, the one-step discrimination—which is a more fine-grained and potentially more sensitive measure—reveals a clear difference between the training conditions. On the whole, the results are again in line with the hypothesis that a change in phonetic category representations underlies the perceptual learning effect.

Signal Detection Analysis

A primary purpose of this study was to test whether the perceptual learning effect influences discrimination performance in the same way that it affects categorization. In particular, if discrimination were not affected by training with the ambiguous sound, then a representational explanation of the categorizaton results would be in doubt. We did, however, find a discrimination effect. What can we infer from this? Unfortunately, although this result is consistent with the representational account, it is ambiguous: A bias change could also explain the discrimination effect (we expand on this below). Therefore, we used signal detection theory to separate the two possible sources of behavioral change.

Before turning to the signal detection analysis, we describe in more detail how changes in representation and bias might be at work in the experimental tasks. For simplicity, we use a straightforward node-activation model (depicted in Figure 6), with frication center frequency as the unidimensional acoustic input. Before training (Figure 6A), the /s/ and /ʃ/ phonetic category nodes are activated by certain ranges of frication center frequencies (ambiguous frequencies are depicted with shading). Activation then feeds into a decision level at which a bias factor is applied, such as that in Luce's choice rule (Luce, 1963). The result is a response label that, in the lexical decision task, determines lexical activation and the resulting word or no word response. In the categorization task, the product of the decision level directly determines the /s/ or /ʃ/ response.

Consider the effect of lexical decision training on this system. One possibility is that it alters the mapping between acoustic–phonetic input and category activation, as shown in Figure 6B for the s-training condition. After training, the range of center frequencies that activates the /s/ node includes the previously ambiguous region, whereas the range that activates the /ʃ/ node no longer includes that region. The result for the categorization task would be a shift in the /s/–/ʃ/ boundary, with more items classified as /s/—particularly in the ambiguous range. Another possible effect of training is to change the bias weights for the two categories, as shown in Figure 6C. In the s-training condition, a bias toward assigning the /s/ label would ease identification of the ambiguous items as words. The mapping between the acoustic–phonetic input and the phonetic category nodes has not changed from pretraining (in Figure 6A). Rather, there is a greater /s/ bias at the decision level, which would result in a categorization boundary shift because less perceptual evidence would be required to give an /s/ label. Of course, a third

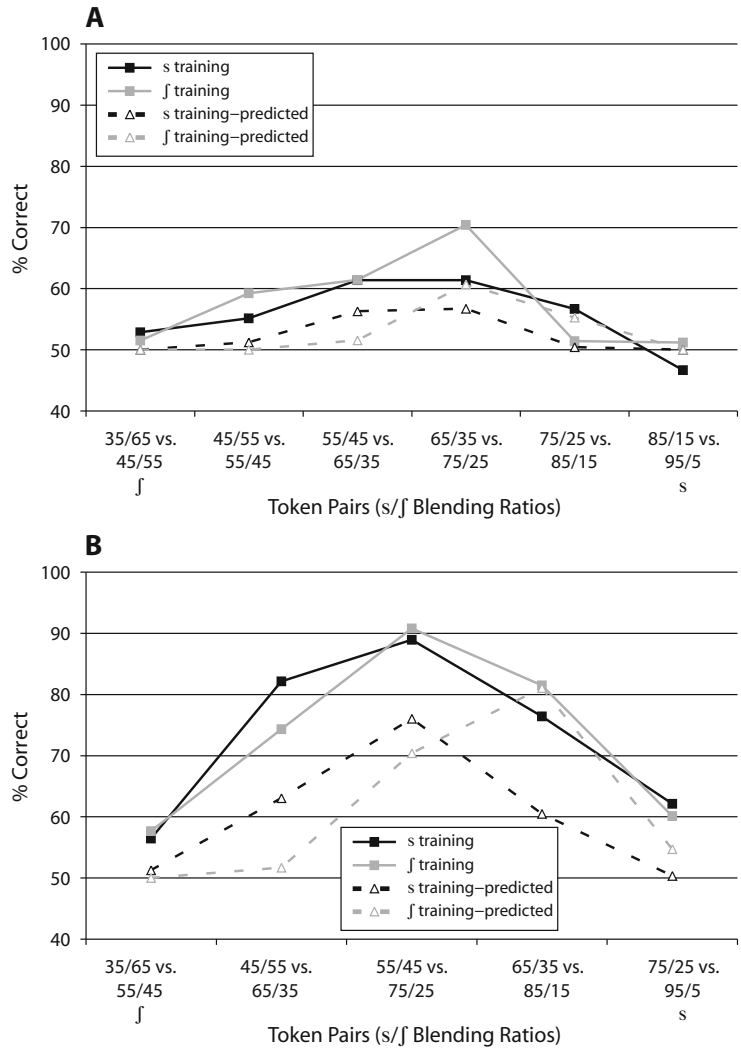

Figure 5. Experiment 2 actual and predicted AXB discrimination accuracy displayed by training condition for (A) one-step and (B) two-step trials. One- and two-step trials were mixed within the discrimination test. Chance is 50%.

possibility is that both a remapping and a bias change occur (not shown).

How would these two alternatives affect the perceptual discrimination task? We assume that discrimination decisions are performed with a phonetic strategy. That is, stimuli are categorized as either /s/ or /ʃ/; then, they are compared (Liberman et al., 1957). This assumption is supported by the presence of discrimination peaks near the categorization boundaries in both experiments. Within this model, the source of the category labels could be either the phonetic category node level or the label level. The remapping hypothesis predicts a difference in discrimination peaks for the two training conditions (no matter which level is used) because the remapping changes phonetic category activation and the resulting labels. However, the bias hypothesis only predicts a peak difference if the label level is used. Therefore, if we had found no difference between the conditions in discrimination, it would have indicated that the categorization results were due primarily to a bias change and that the discrimination performance was based on the phonetic category level (which was the same for the two conditions; see Figure 6C). However, we did find a difference in discrimination performance (for all conditions except two-step discrimination in Experiment 2). This result is consistent with the remapping hypothesis; however, it is also consistent with the bias hypothesis if the biased labels were used to discriminate stimuli. We turned to signal detection analysis to help distinguish between these possibilities.

A signal detection analysis of the group categorization data provides evidence against a bias account of the perceptual learning effect. To preview, the analysis shows

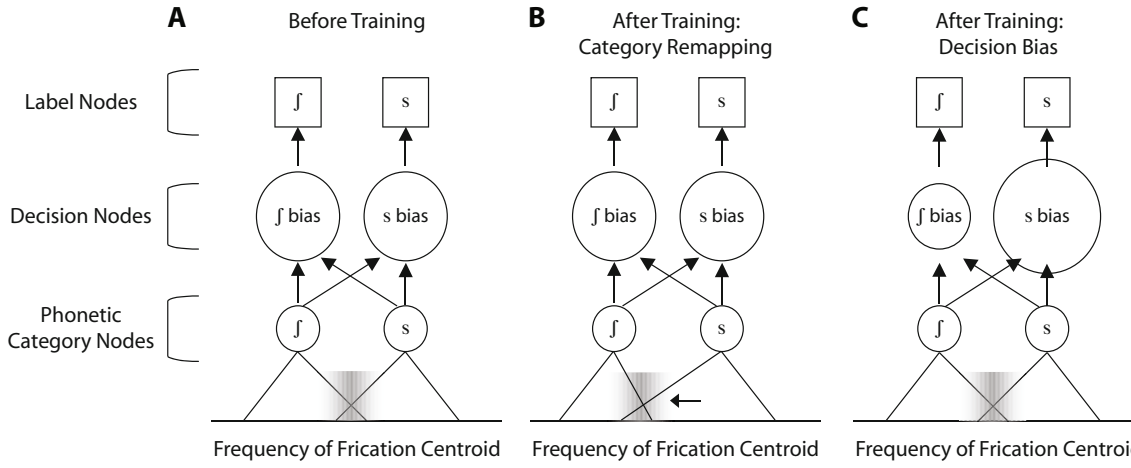

Figure 6. Node activation model depicting possible effects of perceptual training that ambiguous /s/–/ʃ/ sounds are /s/. The system is shown: (A) prior to training, (B) following training, with a change in the mapping between acoustic–phonetic input and phonetic categories, and (C) following training, with an increased bias toward the /s/ label. (Note that the use of bottom-up arrows is simply for expository convenience and is not meant to imply only a bottom-up flow of information within the spoken-word recognition system.)

that the sensitivity parameters differ substantially between the s- and ʃ-training groups in both experiments, and Experiment 2 shows an additional effect of bias. Durlach and Braida (1969; Braida & Durlach, 1972) provided a method for performing signal detection analysis with categorization data for a stimulus continuum, such as that used in these experiments (see also Massaro, 1989; Pitt, 1995; and Sawusch, Nusbaum, & Schwab, 1980, for use of this technique with categorization data). Instead of two probability distributions (signal and noise) as in classical signal detection analysis (Green & Swets, 1966), each stimulus in the continuum (seven in this case) is assumed to have its own evidence probability distribution, and a separate measure of sensitivity (d') is calculated for each adjacent pair of stimuli. If Stimuli X and Y lie next to each other on the continuum with X closer to the /ʃ/ end, then the hit rate is the cumulative percent /ʃ/ response for Stimulus X and all tokens on the /ʃ/ side of it, and the false alarm rate is the cumulative percent /ʃ/ response for Stimulus Y and all tokens on the /s/ side of it. Hit and false alarm rates are calculated in this way for each X–Y pair along the continuum.

Ideally, a separate signal detection analysis would be performed for each participant's data, allowing for statistical tests of the group differences. However, doing so requires each participant to give a large number of responses to each stimulus, and the categorization effect has been found to fade with extensive testing (Kraljic & Samuel, 2006; van Linden & Vroomen, 2007). We therefore limited the test to 10 repetitions of each stimulus token and collapsed the data across participants in order to get stable estimates for a group analysis. Details of the analysis and a discussion of the assumptions underlying its use can be found in the Appendix. The categorization data for Experiments 1 and 2 were submitted to this analysis, with the data collapsed across categorization–discrimination test order for Experiment 1.

The results for Experiment 1 are shown in Figure 7. The patterns of d' values differ for the two training groups, most clearly for the center three stimuli (3, 4, and 5). In terms of cumulative d', Stimuli 3, 4, and 5 for the ʃ-training condition are shifted toward the /ʃ/ end of the continuum by .53, .66, and .36 d' units, respectively, as compared with the s-training condition. In addition, Stimuli 3 and 4 are closer for the ʃ-trained group, whereas Stimuli 4 and 5 are relatively further from each other, indicating that Stimulus 4—the most ambiguous token—was especially affected by training. In contrast to the pattern of d' values, the bias measures did not differ between training conditions. In this analysis, the location of the decision criterion (the measure of bias) is a point on the cumulative d' scale that divides the scale into two regions, corresponding to the two response alternatives. For both conditions, the response criterion was 2.05. According to this analysis for Experiment 1, the difference between the categorization functions for the two training conditions is due entirely to differences in sensitivity.

The analysis for Experiment 2 produced a similar result for sensitivity (not shown). In the ʃ-training condition, Stimuli 3, 4, and 5 shifted toward the /ʃ/ end of the continuum by .69, .66, and .57 d' units, respectively, as compared with the s-training condition. However, the response criterion measures were 2.33 for ʃ training and 2.05 for s training, indicating that—unlike in Experiment 1—there was a response criterion difference in the expected direction. Since the sensitivity change was comparable in Experiments 1 and 2, whereas the effect size of the categorization effect in Experiment 2 was approximately 50% greater than that of Experiment 1 (Experiment 1, categorization difference = 5.54%, $\eta^2_G = .22$; Experiment 2, categorization difference = 10.84%, $\eta^2_G = .33$), the difference in response criteria must account for about one third of the effect.

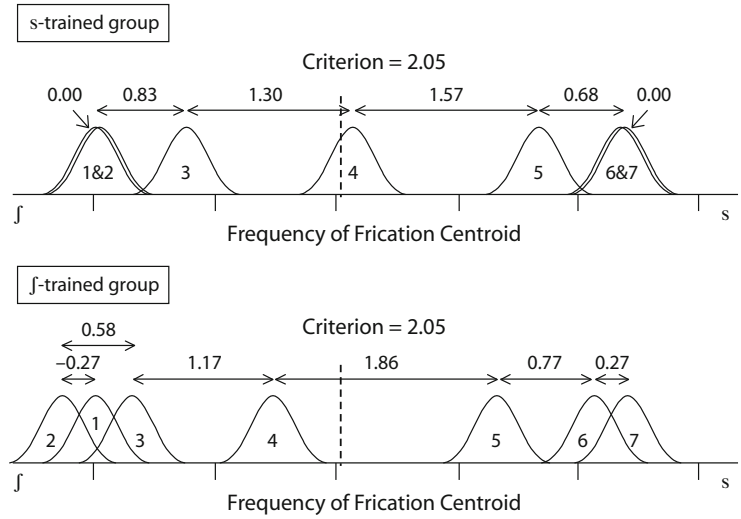

Figure 7. Depiction of response criterion and pairwise *d′* values for the Experiment 1 categorization data for s-training and ʃ-training conditions. Each normal curve represents the evidence (perceptual) distribution for one token on the /s/–/ʃ/ stimulus continuum. Each number indicating the distance between two adjacent stimuli refers to the *d′* value for that pair of stimuli. The vertical dashed line indicates the response criterion on the cumulative *d′* scale.

The fact that training in Experiment 2 induced both category retuning (sensitivity changes) and a bias change may explain the discrepancy between the actual and predicted scores for the two-step discrimination trials (see Figure 5B). The categorization boundaries reflected both the category remapping and the decision bias. In contrast, the discrimination performance—which we propose was based on the phonetic category level (see Figure 6)—only reflected the retuned categories; the group difference was therefore smaller than predicted from the categorization data. This interpretation is also supported by the fact that the categorization effect size was larger in Experiment 2 (η^2_G = .33) than in Experiment 1 (η^2_G = .22), but the discrimination effect sizes were similar (Experiment 1, $\eta^2_{G,1\text{-step}}$ = .05, $\eta^2_{G,2\text{-step}}$ = .03; Experiment 2, $\eta^2_{G,1\text{-step}}$ = .04, $\eta^2_{G,2\text{-step}}$ = .03). Despite the finding of a bias effect in Experiment 2, the presence of large *d′* differences between training conditions indicates that the divergence in categorization performance is due largely to differences in sensitivity. Ironically, we designed this task to be less likely to induce bias; however, bias was found for this task but not for the lexical decision task. Nevertheless, the replication of the perceptual learning effect with a new task is a further demonstration of the robustness of this effect.

GENERAL DISCUSSION

In this study, we examined two possible underlying causes of the perceptual learning effect. In this effect, exposure to a sound that is ambiguous between two phonetic categories in a lexically disambiguating context results in a shifted categorization boundary (Norris et al., 2003). The boundary shift has been credited to the modification of phonetic category representations. However, we tested the possibility that the effect is due to a decision bias resulting from the exposure task typically used to elicit it: lexical decision.

In total, the results of this study suggest that the perceptual learning effect is not due to a decision bias. First, training with the ambiguous sounds not only affected categorization responses, but also influenced the listeners' ability to discriminate tokens on an /s/–/ʃ/ continuum. In Experiments 1 and 2, both the categorization boundary and the peak in discrimination accuracy were further toward the /s/ end of the continuum in the ʃ-training condition than in the s-training condition. This result is expected, given (1) different mappings between acoustic–phonetic space and phonetic category representations following the two training conditions, and (2) the use of phonetic categories in making discrimination decisions. Evidence for the second condition comes from the discrimination peaks themselves. The poor discrimination of pairs within categories and the peaks near category boundaries reveal the strong involvement of category representations. But it is the first condition—the remapping condition—that is of central interest to the present study. As was stated above, if training causes a change in /s/ and /ʃ/ category representations, then discrimination performance is predicted to change. If we had not found different discrimination peak locations depending on training, a phonetic representation account of the perceptual learning effect would be in question.

The second piece of evidence suggesting that the perceptual learning effect is not due to a decision bias is the replication of the effect with the same–different exposure task in Experiment 2. Our reasoning in choosing the same–

different task was that, because it could be performed by simple acoustic comparison of the first and second items, listeners did not have to decide whether the ambiguous sound was an /s/ or an /ʃ/ in order to respond. Therefore, they had little reason to develop a decision bias. This is not to say that the ambiguous sound was not ultimately associated with one of the categories; it must have been, or else the perceptual learning effect would not have occurred. However, we assumed that there was no impetus at the response level to change a decision criterion. As it turned out, this task may have nonetheless induced a decision bias in addition to a representational change, as suggested by the signal detection analysis. We take this finding as a caution against depending on assumptions about how a particular task will influence processing strategies.

The third piece of evidence indicating a change in phonetic category representations came from the signal detection analyses. The categorization performance for both Experiments 1 and 2 showed a substantial shift in the pattern of sensitivity in the predicted directions. In both cases, the boundary stimuli for the ʃ-trained groups were pulled toward the /ʃ/ end of the continuum, as compared with those for the s-trained group. This result suggests a warping of the perceptual space to separate the boundary stimuli from the /s/ end of the continuum when listeners are trained that ambiguous sounds belong to the /ʃ/ category, and from the /ʃ/ end when listeners are trained that they belong to the /s/ category.

The goal of this study was to clarify the source of the perceptual learning effect in speech. We offered the possibility that previous findings of the effect were due to decision biases produced by the training task rather than an underlying modification of phonetic representations. Three key pieces of evidence indicate that bias is not a major component of the perceptual learning effect: (1) the effect of training on phonetic discrimination performance in addition to categorization, (2) the reproduction of the categorization boundary shift with a new training task, and (3) the effect of training on sensitivity for the categorization stimuli, as shown by signal detection analyses. Although a bias difference between groups was discovered in Experiment 2, it was accompanied by a large difference in sensitivity, presumably reflecting group differences in the mapping between acoustic–phonetic space and /s/ and /ʃ/ phonetic category representations.

The present findings fit with previous evidence for specificity and automaticity in perceptual learning. Perceptual learning of the /s/–/f/ contrast (Eisner & McQueen, 2005) and the /s/–/ʃ/ contrast (Kraljic & Samuel, 2005) is tied closely to the acoustic signal in training and does not generalize to stimuli with different acoustic characteristics. A decision bias would presumably produce the effect as long as the task requires the same phonetic labels as those used during training. Therefore, the sum of evidence from the present and previous research points to a representational change in perceptual learning.

Our findings that the perceptual learning effect is a representational phenomenon add to a growing understanding of perceptual learning in speech. Previous studies have shown that the effect is rapid, automatic, acoustically based, driven by lexical feedback (at least when the training stimuli are fully ambiguous), and long-lasting (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007; McQueen, Norris, & Cutler, 2006; Norris et al., 2003). On this foundation, future research can answer the remaining questions about perceptual learning. Most importantly, what are the mechanisms that allow phonetic knowledge to have this highly flexible character while still providing a stable basis for language processing?

AUTHOR NOTE

C.M.C.-D. is now at the Department of Psychology, University of Alberta. This research was funded by a National Institutes of Health National Research Service Award to C.M.C.-D. The authors thank Steve Berg for assistance in data collection. Correspondence concerning this article should be addressed to C. M. Clarke-Davidson, Department of Psychology, P217 Biological Sciences Building, University of Alberta, Edmonton, AB, T6G 2E9 Canada (e-mail: cmclarke@ualberta.ca).

REFERENCES

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, **37**, 379-384.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, **14**, 592-597.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer (Version 4.3.13) [Computer program]. Available from www.praat.org/.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, **51**, 483-502.

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

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, **46**, 372-383.

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

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Krieger.

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

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory & Language*, **56**, 1-15.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence: Brown University Press.

Ladefoged, P. (1989). A note on "Information conveyed by vowels." *University of California Working Papers in Phonetics*, **72**, 161-163.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, **29**, 98-104.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, **55**, 306-353.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.

LUCE, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.

MASSARO, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, **21**, 398-421.

MCGARR, N. S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech & Hearing Research*, **26**, 451-458.

MCQUEEN, J. M., CUTLER, A., & NORRIS, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, **30**, 1113-1126.

MCQUEEN, J. M., NORRIS, D., & CUTLER, A. (2006). The dynamic nature of speech perception. *Language & Speech*, **49**, 101-112.

NEWMAN, R. S., CLOUSE, S. A., & BURNHAM, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, **109**, 1181-1196.

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 & 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.

OLEJNIK, S., & ALGINA, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, **8**, 434-447.

PITT, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1037-1052.

POLLACK, I., & PISONI, D. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychonomic Science*, **24**, 299-300.

SAWUSCH, J. R., NUSBAUM, H. C., & SCHWAB, E. C. (1980). Contextual effects in vowel perception II: Evidence for two processing mechanisms. *Perception & Psychophysics*, **27**, 421-434.

VAN LINDEN, S., & VROOMEN, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception & Performance*, **33**, 1483-1494.

VITEVITCH, M. S., & LUCE, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, **40**, 374-408.

APPENDIX

The signal detection analysis used in the present article is based on a model proposed by Durlach and Braida (1969). Because it is done with group data rather than with individual listener data, and because we cannot assess the extent to which the assumptions in the use of the model are met by our data, this analysis must be treated as exploratory. The purpose of this Appendix is to describe the computation of sensitivity and response criterion in the Durlach and Braida model and the assumptions inherent in its use.

Durlach and Braida (1969) proposed a variant of the classical signal detection theory (Green & Swets, 1966) to account for perception of intensity. Braida and Durlach (1972) extended the basic model to paradigms in which a single stimulus is presented on each trial. The model is designed to deal with continua of N stimuli ($N \geq 2$) and tasks involving M response alternatives ($M \geq 2$). The present experiments involved $N = 7$ stimuli and $M = 2$ response choices (/ʃ/ and /s/). Each of the N stimuli is assumed to give rise to a distribution of internal coding over trials. These evidence or perceptual distributions are assumed to be Gaussian in shape and to have equal variance. From the confusion matrix (the probability with which each of the M responses is used for each of the N stimuli), estimates can be derived of the mean of the probability density function for each stimulus. This results in $N - 1$ values of d' that represent the perceptual distances between adjacent stimuli on the continuum. The location of the $M - 1$ response criteria that separate each of the adjacent response alternatives can also be derived. Massaro (1989) has a concise and easy-to-follow description of this computation.

In our analysis, the group data from all listeners in each condition were used. The response matrix of the probability of each response for each stimulus was converted into a matrix of cumulative probabilities over the response alternatives. That is, for each stimulus, the probabilities for responses 1, 2, 3, . . . , M were replaced by the cumulative probability of using response 1; 1 and 2; 1, 2, and 3; . . . (all of the M responses). The cumulative probabilities were then converted to z scores, with the restriction that only probabilities in the range of .008 to .992 were used, because z scores, with probabilities outside this range would quickly approach infinity. The value of d' for any pair of stimuli was then computed by taking the difference between the z scores for the two stimuli and averaging this across the first $M - 1$ responses (by definition, the cumulative response probability over all M responses is 1.0 for each stimulus). Since there were only two response alternatives in these experiments, only one difference between z scores was computed for each pair of stimuli. In essence, the z scores for each pair of stimuli across the $M - 1$ responses constitute an ROC curve for that pair of stimuli. With only two response alternatives, this amounts to a single point in the ROC space for each pair of adjacent stimuli in the continuum. Finally, the locations of each of the $M - 1$ response criteria were computed for each of the N stimuli and then averaged across the stimuli. Since there were two response alternatives, one response criterion was estimated.

A number of cautionary notes must be considered in using this analysis with the /s/–/ʃ/ continuum data. First, our analysis used group data for each condition rather than individual listener data. Since we had relatively few responses from each listener to each stimulus, individual estimates of d' and response criterion placement would not be reliable. Thus, our analysis is of the aggregate data, and we cannot establish the extent to which the results would be found for each individual listener.

Second, this analysis is appropriate for a perceptual continuum that is unidimensional. Speech distinctions, however, are well known for being multidimensional. In the context of our /s/–/ʃ/ continuum, however, this may not be an issue, since one of the major differences between /ʃ/ and /s/ is in the centroid of the fricative energy distribution (see Newman et al., 2001). Even if the signal is multidimensional, it may be adequate to model it as collapsed onto a single perceptual dimension.

The third caveat is that since our experiments used only two response alternatives and relatively few responses from each listener, we could not assess the degree to which our data met the equal variance and normal distri-

APPENDIX (Continued)

bution assumptions. In an investigation of vowel perception using an /i/–/ɪ/ continuum, Sawusch et al. (1980) collected sufficient data from each listener and found that the equal variance and normal distribution assumptions accounted for 95% of the variance in the data. Of course, just because these assumptions were met for a vowel continuum does not imply that they would be met for our fricative continuum. However, the Sawusch et al. results do show that it is possible to meet the assumptions of this model with speech stimuli. Braida and Durlach (1972) also reported analyses showing that estimates of d' and response criterion were quite robust to violations of these assumptions.

Finally, a signal detection analysis divides the complex processes of perception up into two sets: processes that influence sensitivity and processes that influence the response criterion. In a complex perceptual process such as that involved with speech, this poses a challenge for the interpretation of the signal detection results. In essence, the question is which stages (or processes) influence sensitivity versus which stages (or processes) influence the decision criterion. Two examples will make this question more clear. The sensitivity parameter of the signal detection analyses may reflect only the earliest stages of perceptual processing that involve basic auditory coding. Any complex auditory coding, phonetic categorization processes, and response choice by the listener are represented by effects on the response criterion. Alternatively, all of perceptual processing—up through phonetic coding—may be reflected in the sensitivity parameter, and only the response-choice process influences the placement of the response criterion.

The ambiguity in interpretation represented by these two alternatives is a result of using a one-stage model (signal detection) in the context of a possibly multistage perceptual process. However, in all alternative interpretations, the listener's process of response selection should influence the response criterion (bias), and early coding processes should influence sensitivity. Thus, if the signal detection analysis shows that part or all of a change in the listeners' responses between conditions is due to a change in sensitivity, this implies that the influence of the conditions was (at least partly) on a perceptual process prior to the choice of response. Consequently, a signal detection analysis could provide evidence about the nature of the influences of perceptual learning on perception and help to distinguish between the category retuning explanation and a response bias explanation. The present analysis must be treated cautiously, however, because of all the assumptions made in the use of the signal detection model.

(Manuscript received August 9, 2007;
revision accepted for publication November 14, 2007.)