Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Psychology 51 (2005) 141–178
Cognitive Psychology
www.elsevier.com/locate/cogpsych

Perceptual learning for speech: Is there a return to normal?

Tanya Kraljic *, Arthur G. Samuel

State University of New York at Stony Brook, USA

Accepted 31 May 2005
Available online 10 August 2005

Abstract

Recent work on perceptual learning shows that listeners' phonemic representations dynamically adjust to reflect the speech they hear (Norris, McQueen, & Cutler, 2003). We investigate how the perceptual system makes such adjustments, and what (if anything) causes the representations to return to their pre-perceptual learning settings. Listeners are exposed to a speaker whose pronunciation of a particular sound (either /s/ or /ʃ/) is ambiguous (e.g., halfway between /s/ and /ʃ/). After exposure, participants are tested for perceptual learning on two continua that range from /s/ to /ʃ/, one in the Same voice they heard during exposure, and one in a Different voice. To assess how representations revert to their prior settings, half of Experiment 1's participants were tested immediately after exposure; the other half performed a 25-min silent intervening task. The perceptual learning effect was actually larger after such a delay, indicating that simply allowing time to pass does not cause learning to fade. The remaining experiments investigate different ways that the system might unlearn a person's pronunciations: listeners hear the Same or a Different speaker for 25 min with either: no relevant (i.e., 'good') /s/ or /ʃ/ input (Experiment 2), one of the relevant inputs (Experiment 3), or both relevant inputs (Experiment 4). The results support a view of phonemic representations as dynamic and flexible, and suggest that they interact with both higher- (e.g., lexical) and lower-level (e.g., acoustic) information in important ways.

Keywords: Perceptual learning; Speech perception; Adjustments; Speaker-specificity; Partner effects; Speaker–listener adaptation; Accent learning; Variability; Phonetic adjustments; Memory for voices

* Corresponding author.
E-mail address: tkraljic@hotmail.com (T. Kraljic).

0010-0285/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2005.05.001

1. Introduction

> "You like tomato, and I like tomahto...you say laughter, and I say lawfter" (Ira Gershwin, *Let's Call the Whole Thing Off*)

Despite encountering many different pronunciations of the same sound, some 'normal' and others very atypical (like lisps, for example, or nonnative accents), listeners usually have very little difficulty accurately perceiving a talker's speech, particularly once they have had a bit of experience with that talker. Typically, language researchers attribute such perceptual constancy to a speech perception system that somehow normalizes (i.e., filters) idiosyncratic details of speech. In these theories, the perceptual system essentially overcomes the variability inherent in the speech signal by removing it and reducing the speech signal to a pure, abstract form. This information reduction then allows access to abstractly defined, stable phonemic representations that we maintain in memory, and leads to successful speech perception even in the face of atypical speech (see, e.g., Goldinger, 1998; Tenpenny, 1995, for reviews). In fact, prototypical models of language argue that normalizing the signal is our only hope for perceiving any speech accurately and making sense of what we hear, since virtually no two instantiations of a particular phoneme are identical even *within* a speaker (and certainly not across speakers). This is a view of a very stable perceptual system, one in which the parameters of phonetic categories are set (probably in infancy) and subsequently do not adjust to the countless variations that we come across every day. Instead of dynamically adapting to experiences, as other cognitive systems do, the speech perception system continuously compensates for the signal's endless variability. Even in theories that allow for some learning, its purpose is to make abstraction easier or more efficient (e.g., the learning of "perceptual operations"—Clarke, 2000; Kolers, 1976; Nygaard, Sommers, & Pisoni, 1994).

These theories are almost exclusively based on language processing in isolation or in monologue, and reflect the heavy influence that theoretical linguistics has had on psycholinguistics. But recent research looking at syntactic, lexical, and prosodic choices in dialogue suggests that both conceptual and perceptual representations may be strongly dynamic. People shape each others' linguistic representations and behavior—syntactic, lexical, and prosodic choices often reflect experience with a particular partner (e.g., Branigan, Pickering, & Cleland, 2000; Brennan & Clark, 1996; Garrod & Anderson, 1987; Goldinger, 1998; Levelt & Kelter, 1982). Such adaptations essentially serve to constrain production and comprehension with respect to that partner. Normal conversations have less ambiguity (Brown-Schmidt, Campana, & Tanenhaus, 2005) and less lexical, syntactic, and phonological variability (Brennan, 1999; Coupland, 1984) than language use in isolation has. Speaker–listener convergence thus reduces variability in the signal, but it does so by maintaining flexible representations that reduce the amount of work either person has to do in communication, rather than increasing it.

Some recent work suggests that speech perception may work in much the same way. A growing body of research supports the idea that listeners' perceptual categories, rather than being broadly defined abstractions, actually incorporate

information about particular speakers and contexts (e.g., Goldinger, 1996, 1998; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Schacter & Church, 1992). A consequence of incorporating new instances of speech is that the relevant representations are dynamically adjusted.

In fact, a number of studies have shown that listeners are able to dynamically adjust to speech that is slightly atypical or initially difficult to understand. For example, improved intelligibility has been found after sufficient exposure to synthetic speech (Greenspan, Nusbaum, & Pisoni, 1988; Maye, Aslin, & Tanenhaus, 2003), compressed speech (Dupoux & Green, 1997; Mehler et al., 1993), and nonnative accented speech (Bradlow & Bent, 2003; Bradlow, Pisoni, Akehane-Yamada, & Tohkura, 1997; Bradlow, Akehane-Yamada, Pisoni, & Tohkura, 1999). Clearly, listeners learn something about the speech they are exposed to and use what they learn to improve their perception of similar speech.

Norris, McQueen, and Cutler (2003) demonstrated that listeners are able to adjust (*retune*) their phonetic representations according to the speech they hear with surprisingly little exposure; just 20 of the same 'oddly' pronounced phonemes embedded somewhere in 200 words will lead to perceptual learning. Bertelson, Vroomen, and de Gelder (2003) showed that such learning can also be driven by a visual context that disambiguates the 'odd' phoneme (they call the effect *recalibration*). Maye et al. (2003) found that listening to accented English speech results in phonemic categories that are adjusted in the specific direction of the accent, and not simply in more broad or sloppy categories. And recent work from our own laboratory (Kraljic & Samuel, in press) suggests that retuning for stops occurs at the lowest linguistic level, the feature level (although Eisner & McQueen, 2005, report data which suggests that a locus of adjustment may be at the phoneme level for more acoustically based fricatives). Clearly, then, phonemic boundaries are somewhat flexible and can adapt to variations in the signal, at least for the short term.

But what happens to listeners' phonemic representations *after* they have been adjusted to reflect a talker's speech? The present research focuses on two critical issues related to this question: first, have the representations been modified for the long-term, or simply adjusted temporarily? In other words, do they return over time to their original form, or do they remain in the adjusted form until some further input causes another change? What does the nature of the input have to be to cause the system to return to the original form? There are three logical possibilities: (a) listeners may maintain the adjusted representations until they encounter some linguistic input that "corrects" what they have learned and returns their parameters to the previous (or to a new) setting; (b) listeners may maintain the adjusted representations until they encounter any additional linguistic input, even if it does not contain correcting information about what they previously learned; (c) the representations may simply gradually return over time to their original parameters, even in the absence of any further linguistic input. The latter two possibilities would suggest that perceptual retuning simply reflects a temporary adaptation to the speech we encounter, and not a long-term change in the representations themselves. This would be consistent with a view of a perceptual system that, while able to learn temporarily, maintains stable and abstract long-term phonemic representations.

Of course, if the purpose of perceptual learning is to make communication more efficient, then it would make sense for the system to maintain dynamic representations. But there is a slight complication: we continuously encounter speech from different people, sometimes many times within a single conversation. If we maintain perceptually tuned representations, then what we have learned for one speaker cannot be lost simply because we are now hearing a new speaker. Accordingly, we would expect to find that perceptual learning for a particular speaker persists even after hearing a new speaker with different pronunciations of the critical sounds.

This brings us to the second focus of the present research: are listeners' adaptations specific to the speaker that caused the adaptation, or are they applied more generally to new speakers? For example, do listeners learn: *This odd sound is an /s/*, or do they learn: *This odd sound is an /s/ for Speaker X*? This is another way to address what sort of information makes up phonetic representations. If perceptual learning adjustments are *not* speaker-specific, it would seem that perceptual learning leads listeners to adjust their original phonemic representations; those adjusted representations then are applied generally, without regard to speaker. If, on the other hand, the adjustments *are* speaker-specific, it is certainly possible that the listener is not actually adjusting the category for an /s/, but perhaps instead creating some /s/ 'sub-folder' for that particular speaker (e.g., the listener would have a 'normal' /s/ representation, and a separate 'Speaker X's /s/' representation). In this case, the question that follows is: how is the system able to select the relevant representation for a given speaker? Here, again, there are at least two possibilities: the system may use as a cue the voice of the speaker (i.e., acoustic information), or it may use higher-level information (i.e., knowledge of the speaker's identity). To better understand how the perceptual system adjusts its representations, then, we have conducted a set of experiments that address four specific questions: (1) Given whatever pre-existing phonemic representations a listener has, how do the representations change as a function of new speech input? (2) How do the now-adjusted representations change as a function of subsequent speech input? (3) After any changes have occurred, are they affected by the passage of time (without further speech input)? (4) Are representational changes specific to the voice that provided the new input, or are they more general?

All of the experiments use variations of the perceptual learning paradigm (Eisner & McQueen, 2005; Kraljic & Samuel, in press; Norris et al., 2003). Participants are exposed to ambiguous /s/ or /ʃ/ phonemes in the context of a lexical decision task. Later, they are asked to categorize items on an /s/–/ʃ/ continuum. Differences in the categorization functions as a function of whether the listener heard the ambiguous fricative in an /s/ or an /ʃ/ lexical context assess the presence and extent of perceptual learning.

For clarity of exposition, the presentation of the experiments will be broken into three Parts. Part I has two purposes. First, we establish the perceptual learning effect with our stimuli and procedures. Second, we examine the role that time may play in the modification of perceptual representations. The core of Part I is a comparison of two situations: in one case, listeners undergo the perceptual learning procedure (i.e., the lexical decision task), followed immediately by the assessment of perceptual

learning (i.e., the categorization task). In the second case, the procedure is identical, except that the participants engage in a 25-min silent filler task before the categorization test. Any difference in the size of the perceptual learning effect after this delay reflects the role that time plays in perceptual learning.

Parts II and III are based on a set of three experiments that all share the basic structure of the second case in Part I: participants undergo the perceptual learning lexical decision task, 25 min pass, and perceptual learning is then assessed with the categorization task. Critically, in each experiment in Parts II and III, the 25-min interval comprises an "Unlearning" phase, during which the participants listen to speech input that has the potential to reverse the perceptual learning that has occurred. During Unlearning, participants heard verbal descriptions of pictures that they had to sort into a particular order. The three experiments differ in the nature of the verbal input that was presented during Unlearning. In one case, participants heard descriptions which contained no instances of the critical phonemes (/s/ and /ʃ/). In the second case, the input included unambiguous versions of whichever ambiguous phoneme the listener had heard during the exposure phase. In the final Unlearning situation, listeners heard correct versions of both the ambiguous phoneme they had been exposed to (e.g., /s/) and the other phoneme (e.g., /ʃ/). The three Unlearning conditions test what role, if any, subsequent speech input plays in maintaining or reshifting the perceptual representations of speech.

In Norris et al.'s (2003) seminal study, the authors suggested that a purpose of perceptual learning could be to allow the listener to adjust to the peculiarities of a speaker's accent. If this is indeed the major role of perceptual learning, then such adjustments should be speaker-specific. Two recent studies have examined this issue, with rather different results. Eisner and McQueen (2005) used stimuli similar to those used by Norris et al., based on the fricatives /s/ and /f/. When the exposure phase was based on lexical decisions made to stimuli produced in a female voice, reliable perceptual learning was obtained for tests done in a female voice, but not for tests in a male voice. In contrast, Kraljic and Samuel (in press), using the stop consonants /d/ and /t/, did find reliable cross-voice transfer. In the current study, we continue the investigation of whether perceptual learning is speaker-specific or not, using the kind of same-voice versus different-voice manipulation used by Eisner and McQueen and by Kraljic and Samuel.

Because the experiments in Parts II and III each have three phases (exposure, Unlearning, and categorization), there are two opportunities to explore the speaker-specificity issue: one of these is comparable to what has been tested in these two recent studies—the exposure and categorization tests can either be done in the Same or in Different voices. This manipulation tests whether the perceptual learning is specific to a particular person. We can also manipulate the relationship between the exposure and Unlearning voices. Such a manipulation tests whether returning the phonemic categories to their previous state can be accomplished by any appropriate input, or whether the resetting can only be accomplished by further input from the voice that caused the original perturbation. The exposure-Unlearning speaker specificity issue is examined within each of the experiments in Parts II and III. In Part II, we report the results when perceptual learning is in the same voice as the

categorization test; in Part III, we consider cases in which the speakers in the exposure and categorization tests are different. Collectively, the experiments of the current study provide a comprehensive examination of the conditions that control perceptual learning and relearning.

2. Part I: Do phonemic shifts revert over time, with no further speech input?

2.1. Experiment 1

As we discussed in Section 1, a long-held assumption is that phonemic representations are stable abstractions from all the variations listeners are exposed to. From this perspective, although the perceptual system is able to temporarily learn contextual and speaker related information and use it in comprehension, this information does not become part of the phonemic representation. If this is true, simply allowing some time to pass should cause the learning to fade, and the representations to revert to their pre-perceptual-learning parameters.

Our first experiment tests this hypothesis. Half of the participants in Experiment 1 performed the standard perceptual learning paradigm, with no delay between exposure and test. The remaining participants completed a 25-min, purely visual card sorting task immediately after exposure. During this time, they received no verbal input. After 25 min, participants performed the categorization task, to test whether perceptual learning had faded or still remained.

Our choice of a 25-min delay was based on both practical and empirical grounds. The choice was practical because it resulted in an experimental session that fit within a normal 1-h testing period. The empirical basis for working with a 25-min delay comes from the selective adaptation literature (see Samuel, 1986, for a review). In selective adaptation, categorization shifts are induced through the repeated presentation of an adapting sound. In the large and well-developed literature on adaptation, a study by Harris (1980) suggests that the categorization shifts fade within approximately a half an hour. Thus, even though perceptual learning and selective adaptation are different phenomena (Bertelson et al., 2003; Vroomen, van Linden, de Gelder, & Bertelson, submitted) the adaptation results provide a reasonable starting point for our investigation.

2.1.1. Method

2.1.1.1. Participants. One hundred and ninety-two undergraduate students from the State University of New York at Stony Brook chose to receive either payment or a research credit in a psychology course for their participation in Experiment 1. All participants were 18 years of age or older, and all identified themselves as native English speakers with normal hearing.

2.1.1.2. Design. During the initial Exposure phase, participants performed an auditory lexical decision task that exposed them to a particular speaker's voice (Male or Female) and to an ambiguous phoneme either in /s/ lexical contexts or in /ʃ/ lexical

contexts (?S or ?SH). Participants were randomly assigned to one of the resulting four possible conditions (Male ?S, Male ?SH, Female ?S, Female ?SH).

In the Test phase, participants categorized phonemes that ranged on a continuum from most /s/-like to most /ʃ/-like. All participants categorized /s/–/ʃ/ continua spoken in two different Voices (Same as exposure, Different). The Order of the Same voice (First, Second) was included as a factor for counterbalancing purposes.

In between Exposure and Test, participants in Experiment 1 received either No Unlearning (in which case they completed the Test phase immediately after the Exposure phase, without any delay or additional verbal input) or Visual Unlearning (see below).

2.1.1.3. Materials and procedure

2.1.1.3.1. Phase 1—Lexical decision (Exposure). Two experimental lists were created for an auditory lexical decision task, each with 100 words and 100 nonwords. The lists were identical except for the 40 critical words, as explained in detail in the next section.

Stimulus selection. Forty critical words were selected, ranging in length from two (*brochure*, *obscene*) to four (*negotiate*, *hallucinate*) syllables. None of the critical words contained any instance of the phonemes /z/ or /ʒ/. Twenty of the words also contained no /s/ but each did have a single instance of the critical phoneme /ʃ/. The other 20 critical words contained no /ʃ/; these each instead had a single instance of the phoneme /s/. It was important that the critical phonemes be well articulated and be preceded by reasonably strong lexical information; therefore, the critical words that were chosen had the /s/ or /ʃ/ in the initial position of a syllable that occurred relatively late in the word (e.g., for two-syllable words, the critical phoneme had to appear at the beginning of the second syllable; for three- and four-syllable words, it had to appear at the beginning of the third syllable or later). The two sets of words were matched in mean syllable length as well as in frequency. Critical /s/ words were on average 3.15 syllables in length. Their mean Kucera and Francis (1967) frequency was 17.8, and their mean Zeno, Ivens, Millard, and Duvvuri (1995) frequency was 16.8. Critical /ʃ/ words had 3.15 syllables on average; their mean Kucera and Francis frequency was 22.7 and their mean Zeno et al. frequency was 13.

The 100 filler words were also selected to have no instance of /s/, /ʃ/, /z/, or /ʒ/ phonemes anywhere. The fillers were matched to the critical words in term of stress pattern, number of syllables, and word frequency. Fillers had an average 3.02 syllables and a mean frequency of 13.4 (Kucera & Francis, 1967) and 11.6 (Zeno et al., 1995).

Finally, to ensure equal numbers of 'Word' and 'Nonword' responses in the lexical decision task, we created a nonword for each filler word. We created nonwords by changing one phoneme per syllable of each word; phonemes were changed to another phoneme with the same manner of articulation (i.e., glides changed to glides, stops to stops, etc.). Using this method, we made 100 filler nonwords with no /s/ or /ʃ/. As with all of the words, no /z/ or /ʒ/ appeared in any position either. In order for each list to contain equal numbers of words and nonwords (100 of each), all 100 of the nonwords were used, and 60 of the filler words. Appendix A lists all of the critical words, fillers, and nonwords used in the experiments.

Stimulus construction. Each of the 40 critical words, 60 filler words, and 100 filler nonwords was recorded by both a male and a female speaker. Both speakers had a New York dialect; our male speaker had an average f0 of 120 Hz, and the female speaker's average f0 was 230. Speakers read the words into a microphone in a sound-proof recording chamber, and were recorded at 16 kHz onto a PC using Goldwave sound editing software. Each word and nonword was saved in its own file, and each file was edited to eliminate any background noise.

In addition, each speaker read aloud a second version of the 40 critical words. In this version, the critical phoneme (/s/ or /ʃ/) that normally appeared in the word was replaced with the other one, creating pairs of critical items; for example, each speaker would record both '*brochure*' and '*brosure,*' '*hallucinate,*' and '*hallushinate.*' Recording both an /s/ and /ʃ/ version of the same word allowed us to create a unique ambiguous (/?sʃ/) mixture for each critical word. This was used during the Exposure phase to replace the /s/ in the ?S conditions, and the /ʃ/ in the ?SH conditions (as opposed to creating a single generic /?sʃ/ mixture that could be inserted into every word). Using word pairs preserves any coarticulation information into and out of the critical phoneme, resulting in more natural-sounding items that do not highlight the presence of an ambiguous sound.

The acoustic properties of /s/ and /ʃ/ allowed us to construct the ambiguous /?sʃ/ sound with a relatively simple method. All stimuli were constructed using the /s/-version of the word as the "frame." Because /s/ and /ʃ/ are generally very similar in both duration and amplitude, we were able to mix the two sounds in a straightforward way. The duration of the /s/ was measured first; it was then cut out of the word and saved into its own file. An equal duration of the /ʃ/ was then excised from its version. The /s/ and /ʃ/ were then mixed together with five different weightings that varied from 30% /s/ and 70% /ʃ/ to 70% /s/ and 30% /ʃ/. Each author listened to all five mixtures and independently judged which was most ambiguous for each item; if the authors disagreed by more than one step, the midpoint was used as the most ambiguous. If the authors disagreed by one step, a mixture was created that was midway between the two points. In this way, a single ambiguous mixture for each critical item was selected for use in the experiment. Appendix B lists the mixtures that were used for each item.

Finally, two experimental lists were created so that one experimental group heard words with intact /s/s and ambiguous /ʃ/s (?SH); the other group heard intact /ʃ/-words and ambiguous /s/s (?S). Each list included 40 critical words (20 of which included an ambiguous /sʃ/). Participants in the lexical decision task were randomly assigned to one of the four groups created by crossing the mispronounced phonemes and the two speakers. Up to three participants were tested simultaneously in a sound proof booth. Stimuli were presented over headphones and participants responded 'Word' or 'Non-word' by pressing the corresponding button on a response panel; responses and reaction times were recorded and saved to individual subject files. The experimenter stressed both speed and accuracy in her instructions. She also told the participants that the lexical decision task would be followed by a phoneme categorization task, and she explained each task in detail. The reason for giving all of the instructions initially was to avoid any talking in between Exposure and Test, and

the experimenter stressed this point in her instructions. However, the participants were not told what sound they would be categorizing in the categorization phase, only that the labels on the response panel would be changed accordingly when the time came. They were also not told that some of the words would have ambiguous sounds.

Items were presented in a new random order for each testing session. The presentation of the items was self-paced; a new item was presented 1 s after all participants had responded to the previous item. However, if the participant(s) failed to respond within 4 s, the next item was presented.

2.1.1.3.2. Phase II—Unlearning. As explained in the Design, after doing the lexical decision task, participants were randomly assigned either to a No Unlearning condition or to Visual Unlearning.

Participants who were assigned to the No Unlearning condition performed the category identification task (see below) immediately after the lexical decision task, with no intervening delay.

For the Visual Unlearning, 20 index cards, each depicting a different abstract geometric shape known as a tangram, were placed on a display in front of the participant (see Fig. 1 for examples of these shapes). There were five grey, five yellow, five orange, and five green tangrams. The cards were arranged so that all 20 pictures were visible. The participants had to put all 20 cards into a particular (random) order. The colored shapes on the cards were visually presented in rapid succession on a computer screen in front of each participant. The participants' task was to sort as many of the cards as they could during the visual presentation; after all the shapes had been presented, participants pressed a button and the same random order was presented again, while the participants tried to put a few more cards in order. Each shape was presented for 1 s, followed by 250 ms of a blank screen, and then the next shape. Each random order was presented seven times (first two decks of cards), six times (following two decks of cards), and five times (for all subsequent decks of

Experiment 2 *Neutral Description*	a yellow bridge over a highway	a green one, kind of like a maple leaf	a grey one with a flat line below a white triangle
Experiments 3 & 4* *Corrected /s/*	a yellow rectangular figure with two SQUARES cut out	a green CASTLE with ITS drawbridge down	a grey one that LOOKS like a SAILboat
Corrected /ʃ/	a yellow one SHAPED like a SHORT bridge over a highway	a green one like a road leading to a MANSION	a grey one SHAPED like a SHIP with a thick bottom line

*As described in the paper, participants in Experiment 3 received either the corrected /s/ or the corrected /ʃ/ description for *all* tangrams; participants in Experiment 4 heard /s/ descriptions for half of the tangrams, and /ʃ/ descriptions for the other half.

Fig. 1. Example of several tangrams used in the Unlearning phase, along with descriptions used in Experiments 2, 3, and 4. Words that contain the critical phonemes (/s/ or /ʃ/) are presented in capital letters.

cards). Participants ordered as many decks as they could within 25 min (typically four or five).

No responses or response times were recorded from the Unlearning phase, since the purpose of the task was simply to fill 25 min while ensuring that participants were not getting any speech input.

Phase III—Category identification. In the final phase of the experiment, participants categorized six tokens on two separate /s/–/ʃ/ continua, one in the original voice they had been exposed to (during lexical decision) and one in a new, previously unheard voice. The continua were blocked by voice (Same versus Different). The order of presentation voice was counterbalanced across subjects. The procedure for creating the continuum was the same as that for creating the ambiguous critical items used in the lexical decision task: each of the endpoints of the continua (/asi/ and /aʃi/) were recorded by the same male and female speakers who produced the lexical decision stimuli, and the endpoints were mixed together in proportions varying from 20%s/80%ʃ to the reverse. Six points for each continuum were chosen, ranging in equal steps from relatively /s/-like to relatively /ʃ/-like, with four ambiguous points in between.

At the beginning of the categorization phase, the experimenter changed the labels on each participant's response panel so that one button was labeled "S" and the other was labeled "SH." Participants had previously been told that in the final part of the experiment they would hear vowel–consonant–vowel syllables, and that they should respond as quickly as possible to each syllable, pressing the button on the response panel that corresponded to the consonant they had heard. Ten randomizations of the six sounds on the /s/–/ʃ/ continuum were presented in one voice; then participants performed the same categorization task on the /s/–/ʃ/ continuum in the other voice. Responses and response times were recorded.

2.1.2. Results and discussion

2.1.2.1. Lexical decision. We examined performance on the lexical decision task first, since this would tell us whether our critical (ambiguous) /?sʃ/ items had indeed been perceived by our participants as words. Any participant whose accuracy in categorizing either the critical or the filler items was below 75% was replaced. Twelve of the 192 participants were replaced for this reason.

A summary of the accuracy and RT data for each type of critical item (ambiguous ?S or ?SH versus natural /s/ or /ʃ/) is included in Table 1. Overall, listeners performed very well on the lexical decision task; mean accuracy (for all items) was 96.5%. Participants were slightly more accurate in judging the natural versions of our critical items as words (98.7%) than in judging the ambiguous versions (94.3%), $F1(1,184) = 113.1$, $p < .001$; $F2(1,19) = 15.61$, $p = .001$. There were no differences in reaction time: people were just as fast to categorize the ambiguous versions as words (934 ms) as they were to categorize the natural versions (929 ms), F1 $(1,184) = .454$, $p = .401$; $F2(1,19) = .226$, $p = .64$. Overall, the accuracy and response time data suggest that our ambiguous mixtures were relatively natural sounding.

Table 1
Experiment 1, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?ssh/	
	/s/	/ʃ/	?S	?SH
% Correct	98.7 (2.8)	98.7 (2.5)	93.2 (6.3)	95.3 (5.1)
RT (in ms)	924 (129)	934 (125)	938 (132)	928 (141)

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words. Standard deviations appear in parentheses.

2.1.2.2. Category identification. The purpose of our category identification task is to assess whether listeners have perceptually learned; that is, whether listeners who were exposed to ambiguous /ʃ/ (e.g., in words like bro?ure) would learn to perceive more items on an /asa/—/aʃa/ continuum as an /ʃ/, while those who heard ambiguous /s/ (e.g., hallu?inate) would learn to perceive the same sounds as /s/. Such learning should result in identification functions that are shifted in opposite directions, and functions that are therefore quite different for the two lexical decision groups. For each experiment, we present the size of this difference (calculated as the percentage of /ʃ/ responses for the ?SH lexical decision group, minus the percentage of /ʃ/ responses for the ?S lexical decision group) in the caption to each figure.

Recall that for the present section, we *only* consider cases in which participants are tested on the *same* voice that they were trained on (the Male Exposure–Male Test and Female Exposure–Female Test cases). We do this to remain focused on our primary question: what happens to representations after they have been adjusted, and what does that tell us about the nature of phonemic representations? Results that bear on our second question (Do the adjustments generalize to different speakers), in the Male Exposure–Female Test and Female Exposure–Male Test conditions, will be addressed in Part III.

2.1.2.2.1. No Unlearning Group. There was a clear effect of lexical decision exposure condition on phonemic categorization performance in our No Unlearning Group. People who were exposed to the ambiguous phoneme in words that normally have an /ʃ/ categorized more items on our continua as /ʃ/ (64.6%) than people who were exposed to the ambiguous sound in words which normally have an /s/ (54.4%), $F(1,92) = 8.61$, $p < .005$. Fig. 2 shows the perceptual learning effect for both the Male–Male and Female–Female No Unlearning groups. This overall training effect confirms that people do adjust their perceptual categories of /s/ and /ʃ/ to reflect the speech they are exposed to, and that this adjustment is evident immediately after such exposure. There was no interaction between training condition and voice ($F = .049$, $p = .826$): the perceptual learning effect was reliable for participants who were trained and tested on the Male voice ($F1(1,46) = 4.43$, $p < .05$), just as it was for participants who were trained and tested on the Female voice ($F(1,46) = 4.26$, $p < .05$).

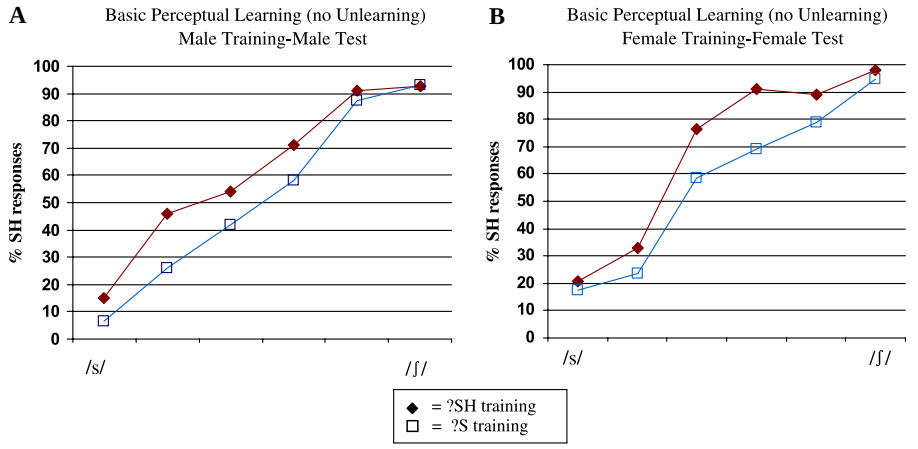

Fig. 2. Experiment 1. (A) The perceptual learning effect for participants trained and tested on the Male voice, with no intervening unlearning condition, reflects a 9.5% difference in responses to /s/ and /ʃ/; and (B) for those trained and tested on the Female voice, no unlearning condition, the shift is 11.1%.

2.1.2.2.2. Visual Unlearning Group. Our main interest in Experiment 1 was to see what would happen to the perceptual learning effect after some time passes: would the perceptual learning effect fade (or disappear altogether)? Fig. 3 shows the perceptual learning effects for both Visual Unlearning groups. The data indicate that perceptual learning does not attenuate with time. After 25 min, participants who had been exposed to ambiguous /ʃ/ categorized significantly more items on our continua

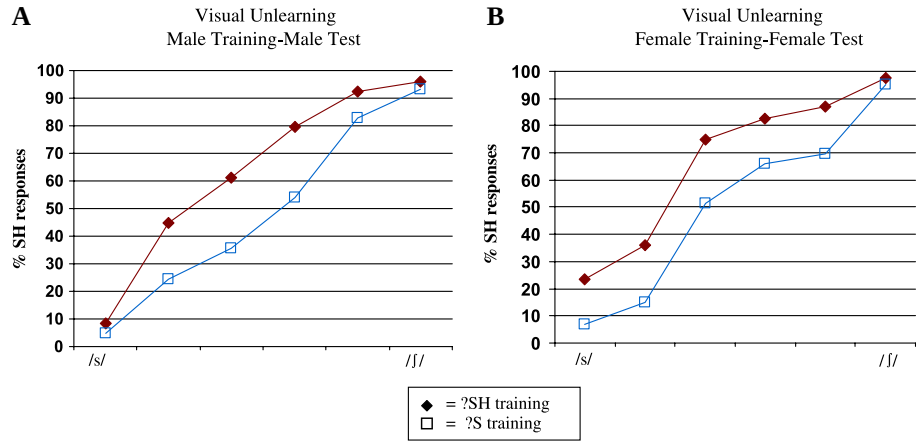

Fig. 3. Experiment 1. The perceptual learning effect is larger after 25 min: (A) a 14.6% shift for participants who trained and tested on the Male voice, Visual unlearning condition; (B) 16.2% for those who trained and tested on the Female voice, Visual unlearning condition.

as /ʃ/ (65.0%) than participants who had been exposed to /s/ (49.7%) ($F(1,46) = 19.7$, $p < .001$). Just as we saw in the No Unlearning condition, this effect is significant for both the Male–Male condition ($F(1,46) = 10.07$, $p = .003$) as well as the Female–Female condition ($F(1,46) = 9.74$, $p = .003$).

The data from Experiment 1 show an intriguing pattern. First, as expected, it is clear that listeners can and do learn to adjust their phonemic categories for /s/ and /ʃ/; participants who were tested immediately after exposure (No Unlearning) showed perceptual learning shifts of approximately 10%. However, rather than this effect fading over time, participants who performed an intervening task for 25 min between exposure and test showed shifts of 15.3%, a numerically larger effect (although not significantly larger: $F(1,184) = 1.075$, $p = .3$). Clearly, the perceptual learning effect does not diminish with time; if anything, it appears to stabilize after some time has passed. This result suggests that perceptual learning does not reflect a transient adjustment of the system, but instead that whatever is learned during perceptual learning becomes part of the phonemic representations.

If this is the case, the adjusted representations should not return to their pre-learning parameters until some further input corrects the adjustment that has been made. But what does the nature of the input have to be to cause the representations to be changed again, back to pre-learning settings? The next three experiments address this question. The first of these experiments tests whether simply hearing 'normal' speech from the same speaker, without any instances of /s/ or /ʃ/ (good or bad), will cause the system to return to pre-learning parameters. It also looks at whether normal speech from *any* speaker (again, with no /s/ or /ʃ/) will have the same effect.

3. Part II: What kind of speech input can reset the perturbed phonemic categories?

3.1. Experiment 2

Experiment 1 demonstrated that perceptual learning effects may actually stabilize as time passes. They are at least as large after 25 min have passed as they are immediately after the system has been exposed to speech. This pattern suggests that the system maintains the adjusted representations until some further input 'corrects' what it has learned. Because Experiment 1's Unlearning condition was purely visual, it is conceivable (although extremely unlikely) that participants engaged in some kind of rehearsal between Exposure and Test, causing the perceptual learning effect to be maintained unnaturally.

This possibility is eliminated in Experiment 2. Experiment 2 provides a verbal equivalent to the visual Unlearning condition: participants hear descriptions of the cards they are sorting, with the descriptions worded to be without any /s/ or /ʃ/ sounds. These descriptions test one of the ways that the system could return to its pre-learning parameters. If the system has learned something about a person's pronunciation of /s/ or /ʃ/ that has now become part of its long-term representation, then hearing that person's (or *any* person's) voice for 25 min should not be enough

to re-set the parameters; the system should require very specific input that directly corrects what has previously been learned.

3.1.1. Method

3.1.1.1. Participants. One hundred and twenty-eight undergraduate students from the State University of New York at Stony Brook chose to receive either payment or a research credit in a psychology course for their participation in this experiment. All participants were 18 years of age or older, and all identified themselves as native English speakers with normal hearing. None had participated in Experiment 1.

3.1.1.2. Materials and procedure. Participants performed the same initial Exposure task (Phase I) and the same final Test (Phase III) as in Experiment 1. In between exposure and test, participants performed a verbal Neutral Unlearning task (Phase II). The materials and procedure for Phases I (Exposure) and III (Test) were identical to Experiment 1.

3.1.1.2.1. Phase II—Unlearning. After doing the lexical decision exposure task, participants performed a Neutral Unlearning task in either the Same voice they had been exposed to in Phase I, or in a Different Voice. The task was to sort the same 20 tangram-cards used in Experiment 1 into a particular (random) order. Participants heard verbal descriptions of the cards over headphones; after each description, they had to pick the correct card out of the display, place it face down in front of them, and press a button to indicate that they had completed that card. Once all the participants in a session had pressed a button, they heard the next description. When all 20 descriptions had been heard, the experimenter took each participant's ordered cards. Participants repeated this task four or five times, until 20–25 min had passed.

The words heard during the Neutral Unlearning described each tangram using no /s/ or /ʃ/ sounds. An example of the Neutral Unlearning description for two of the tangrams can be found in Fig. 1 (along with nonneutral descriptions used in Experiments 3 and 4).

3.1.2. Results and discussion

3.1.2.1. Lexical decision. As in Experiment 1, we examined performance on the lexical decision task first and replaced any participant with lower than 75% accuracy at categorizing either the critical or filler items. Five of the 128 participants were replaced for this reason.

Once again, all participants performed very well in the lexical decision task: mean accuracy overall was 97%. A summary of the accuracy and RT data for each type of critical item (ambiguous ?S or ?SH versus natural /s/ or /ʃ/) is included in Table 2. Participants were slightly more accurate in judging the natural versions of our critical items as words than in judging ambiguous versions ($F1(1,120) = 64.4$, $p < .001$; $F2(1,19) = 21.7$, $p < .001$). They were equally fast to make judgments to ambiguous and natural items ($F1(1,120) = .372$, $p = .534$; $F2(1,19) = .202$, $p = .658$). These data indicate again that our ambiguous mixtures were relatively natural sounding.

Table 2
Experiment 2, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?ssh/	
	/s/	/ʃ/	?S	?SH
% Correct	99.4 (1.7)	98.8 (2.9)	93.8 (7.4)	96 (5.2)
RT (in ms)	927 (108)	935 (107)	937 (114)	934 (127)

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words. Standard deviations appear in parentheses.

3.1.2.2. Category identification. Recall that the purpose of Experiment 2 was to provide a verbal equivalent to Experiment 1's Visual Unlearning condition. Participants in this experiment heard verbal descriptions of the cards they were sorting; the descriptions were presented either in the Same voice as training and test, or in a Different voice. Again, all results reported in this section of the paper will be from those cases in which participants were tested on the same voice that they were trained on.

3.1.2.2.1. Neutral input, same voice condition. Fig. 4 presents the categorization functions for listeners who heard descriptions in the Same voice used in the Exposure phase. Overall, participants who heard descriptions of the tangrams in the same voice as training still showed a significant, and strong, perceptual learning effect ($F(1,60) = 13.588$, $p < .001$). There was no significant interaction with training Voice ($F(1,60) = .184$, $p = .67$). Indeed, the perceptual learning effects for Male–Male and Female–Female are virtually the same size (15.1 and 14.4%, respectively), and both

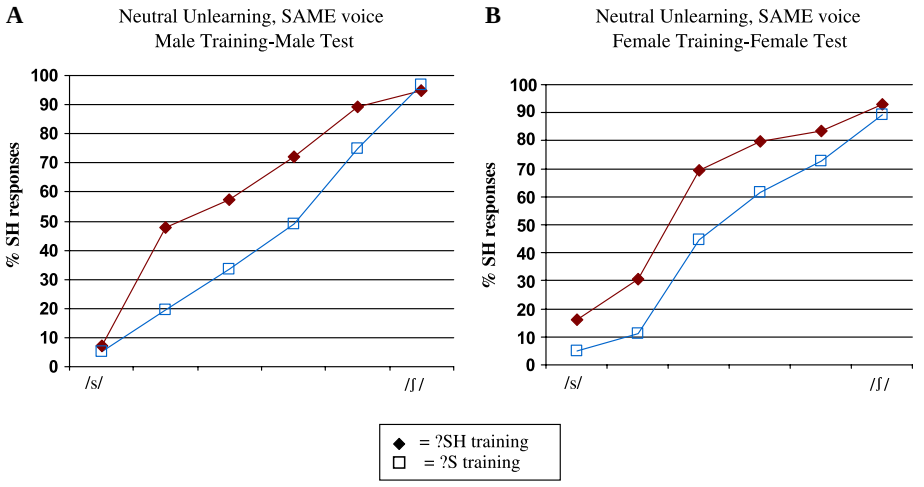

Fig. 4. Experiment 2. Neutral input during Unlearning, in the Same voice as training. The perceptual learning effect is still large for both groups: (A) 15.1% for those who trained and tested on the Male voice and (B) 14.8% for those who trained and tested on the Female voice.

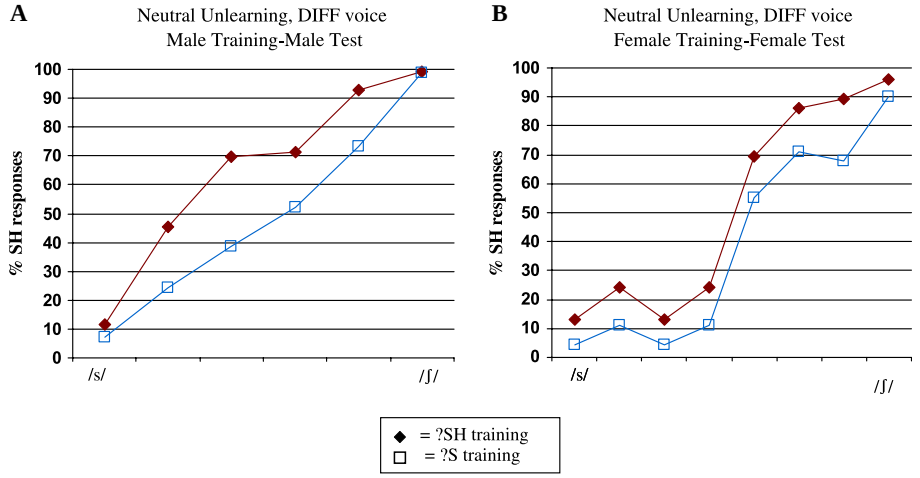

Fig. 5. Experiment 2. Neutral input during Unlearning, in a Different voice from training. The effect is still large for both groups: (A) 16.5% for those who trained and tested on the Male voice and (B) 13.1% for those who trained and tested on the Female voice.

are significant (Male–Male: $F(1,30) = 7.17$, $p < .012$; Female–Female: $F(1,30) = 5.29$, $p = .029$).

3.1.2.2.2. Neutral input, different voice condition. The data from the different voice condition show the same pattern (see Fig. 5). There is a strong overall perceptual learning effect ($F(1,60) = 12.21$, $p = .001$), that is significant for both the Male–Male case ($F(1,30) = 7.19$, $p = .012$) and the Female–Female case ($F(1,30) = 6.45$, $p = .017$).

The results from Experiment 2 confirm those from Experiment 1: perceptual learning effects appear to stabilize after some time has passed, and as a result the effects after 25 min are substantial (approximately 15%, the same size shift that the Visual Unlearning condition produced in Experiment 1). Clearly, being exposed to speech during this stabilization time, even from the speaker whose voice induced the perceptual learning, does not attenuate the boundary shift. However, the speech in Experiment 2 intentionally did not contain any input that explicitly corrected what had been learned: it did not contain any 'good' tokens of /s/ or /ʃ/.

Experiment 3 examines whether hearing such corrected input does, in fact, reset the system to pre-learning parameters. If such correction does take place, a central question is whether this input has to come from the same speaker who had caused the parameters to shift, or whether 'good' tokens from any speaker are sufficient.

3.2. Experiment 3

Experiments 1 and 2 provide clear evidence that perceptual learning is not a transient adjustment to current speech input; rather, whatever is learned is

maintained over the longer term in phonemic representations. If this is the case, the phonemic representations should only return to pre-learning parameters when a listener is exposed to input that corrects what has been learned. Specifically, listeners who have been trained on odd /s/ would need to hear good /s/ to 'undo' the learning, and listeners who have been trained on odd /ʃ/ would need to hear good /ʃ/.

Further, if listeners are learning something about the speaker that is also integrated into phonemic representations, then only correcting information from the *same* speaker as exposure should diminish or eliminate perceptual learning for that speaker.

3.2.1. Method

3.2.1.1. Participants. One hundred and twenty-eight undergraduate students from the State University of New York at Stony Brook chose to receive either payment or a research credit in a psychology course for their participation in this experiment. All participants were 18 years of age or older, and all identified themselves as native English speakers with normal hearing. None had participated in Experiment 1 or 2.

3.2.1.2. Design. Participants performed the same initial Exposure phase and the same final Test phase as in Experiments 1 and 2. In between exposure and test, participants performed a verbal Corrected-phoneme Unlearning task, in which they heard descriptions in either the Same voice as exposure, or a Different voice.

3.2.1.3. Materials and procedure. The materials and procedure for Phases I (Exposure) and III (Test) were identical to Experiments 1 and 2.

3.2.1.3.1. Phase II—Unlearning. The Unlearning task here was the same card-sorting task used in Experiment 2. The single difference was the wording of the descriptions heard by the participants. The descriptions for the Corrected-phoneme Unlearning task included normal versions of the previously mispronounced phoneme. Specifically, participants who heard odd /s/ in the Exposure phase now heard an average of one good /s/ in each tangram's description during Unlearning (but heard no /ʃ/); similarly, participants who heard odd /ʃ/ during Exposure now heard an average of one good /ʃ/ in each tangram's description (but no /s/). See Fig. 1 for examples of the descriptions used in this condition. The procedure for sorting the tangram cards was identical to Experiment 2. Note that because there were 20 cards, and participants typically completed 4–5 decks in the 25 min, about 80–100 corrected phonemes were heard during the Unlearning phase. Participants were randomly assigned to hear the Corrected descriptions in either the Same voice they had heard during Exposure, or in a Different voice.

3.2.2. Results and discussion

3.2.2.1. Lexical decision. Again, we examined performance on the lexical decision task first and replaced any participant with lower than 75% accuracy at categorizing either the critical or filler items. Four of the 128 participants were replaced for this reason.

Participants in Experiment 3 (like those in Experiments 1 and 2) performed very well on the lexical decision task. A summary of the accuracy and RT data for each type of critical item (ambiguous ?S or ?SH versus natural /s/ or /ʃ/) is included in Table 3. Overall accuracy was 97.3%, with ambiguous items being responded to only slightly less accurately than natural items ($F1(1,120) = 67.4$, $p < .001$; $F2(1,19) = 17.45$, $p = .001$), and equally quickly ($F1(1,120) = 1.02$, $p = .314$; $F2(1,19) = .468$, $p = .502$). These data confirm what we have seen in the previous experiments: participants are highly accurate and equally fast at judging the lexical status of our ambiguous and natural words, indicating that our ambiguous items were relatively natural sounding.

3.2.2.2. Category identification. Recall that participants in this experiment all heard descriptions during Unlearning that provided correct (i.e., natural) versions of whichever ambiguous phoneme they had been exposed to during lexical decision. Our critical comparison is between those who heard these corrected phonemes in the Same voice as exposure versus those who heard them in a Different voice. If the system is adjusting parameters in a voice-specific way, we expect to find that additional information from the Same voice affects this adjustment differently than additional information from a Different voice does. In fact, even though there was an overall effect of training condition (57.1% /ʃ/ after /?ʃ/ Exposure versus 49.6% after /s/; $F(1,120) = 6.36$, $p = .013$), the training effect interacted strongly with Unlearning voice ($F(1,120) = 5.743$, $p = .018$).

3.2.2.2.1. Corrected phoneme, same voice condition. There was no significant perceptual learning effect for participants who heard corrected phonemes during Unlearning in the Same voice that they had heard during exposure ($F(1,60) = .007$, $p = .934$). The loss of perceptual learning was observed for both Male training–Male test ($F(1,30) = 1.060$, $p = .312$) and Female–Female ($F(1,30) = .307$, $p = .584$). Fig. 6 shows graphs for these two cases.

3.2.2.2.2. Corrected phoneme, different voice condition. As Fig. 7 shows, perceptual learning was only significant for participants who heard a Different voice at unlearning ($F(1,60) = 14.029$, $p < .001$). Both the Male training–Male test and Female training–Female test cases show robust perceptual learning effects ($F(1,30) = 4.53$, $p = .04$ and $F(1,30) = 9.309$, $p = .005$, respectively). Although the perceptual

Table 3
Experiment 3, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?ssh/	
	/s/	/ʃ/	?S	?SH
% Correct	99.4 (1.9)	99.4 (1.9)	94.8 (6.7)	95.6 (4.1)
RT (in ms)	945 (117)	946 (109)	973 (131)	933 (118)

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words. Standard deviations appear in parentheses.

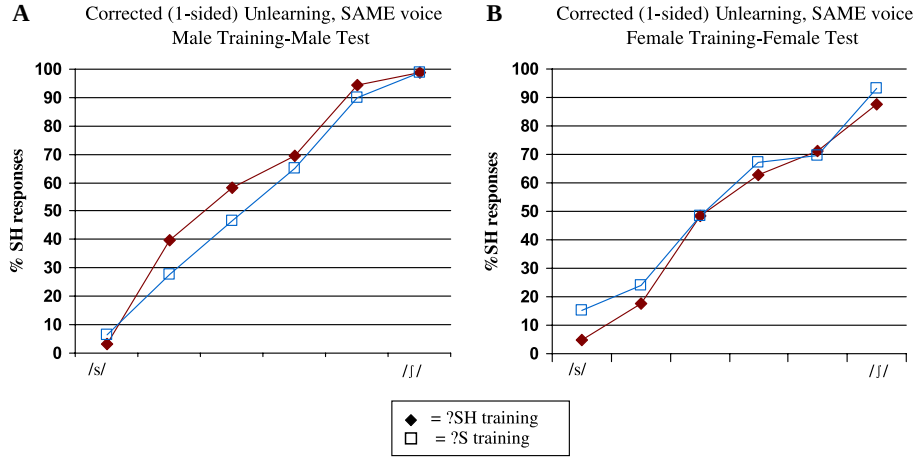

Fig. 6. Experiment 3. Corrected phonemes during Unlearning, in the Same voice as training. The perceptual learning effect was significantly attenuated for both groups: (A) only a 5% shift for those who trained and tested on the Male voice and (B) −4.2% (that is, 4.2% in the wrong direction) for those who trained and tested on the Female voice.

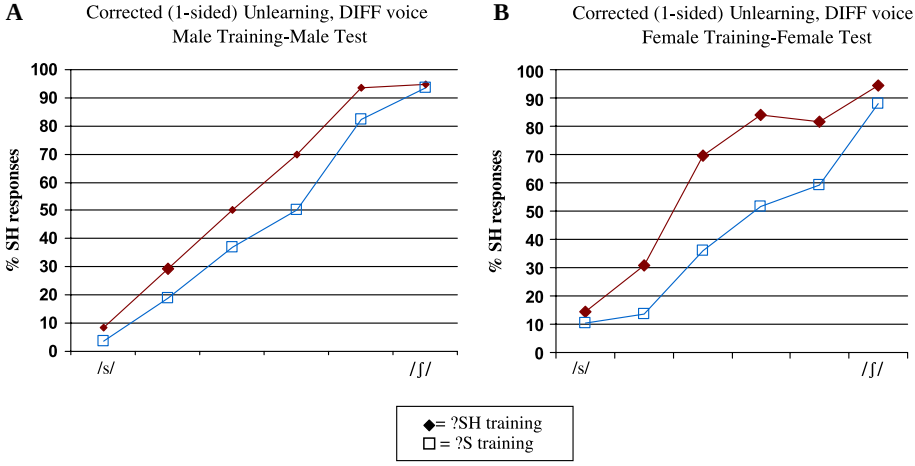

Fig. 7. Experiment 3. Corrected phonemes during Unlearning, in a Different voice from training. The perceptual learning effect is significant for both groups: (A) 10% for those who trained and tested on the Male voice and (B) 19.3% for those who trained and tested on the Female voice.

learning effect for the Female–Female case appears larger than that for the Male–Male case (a shift of approximately 20% versus 10%), this difference did not produce a significant interaction ($F(1,60) = 1.402$, $p = .241$).

The data from Experiment 3 show an intriguing pattern: once participants have adjusted to a particular voice, hearing new (correcting) input from a different speaker

does not affect the adjustment—in fact, as in Experiments 1 and 2, the effect was numerically bigger after the intervening period. However, when the new (correcting) input comes from the *same* speaker, the perceptual learning effect disappears. These results suggest that perceptual learning adjustments are maintained in a speaker-specific way.

However, it is possible that these results might not reflect pure perceptual learning effects. Results of numerous adaptation studies demonstrate that participants who are exposed repeatedly to a particular sound (e.g., /s/ or /ʃ/) subsequently categorize fewer items on a continuum (say, from /s/ to /ʃ/) as that sound. In the current experiment, such an adaptation effect would lead to a reduction in, or possibly elimination of, the perceptual learning effect we are looking at—precisely what we were looking for in Experiment 3. Because it is impossible to tell to what extent the data from Experiment 3 reflect such an adaptation effect versus perceptual (un)learning, we devised a procedure that eliminates the potential influence of adaptation during verbal Unlearning.

3.3. Experiment 4

Experiment 4 measures the perceptual learning effect that remains after listeners are exposed to corrected versions of the ambiguous phoneme they were trained on, under conditions that should eliminate any potential influence of adaptation.

3.3.1. Method

3.3.1.1. Participants. One hundred and twenty-eight undergraduate students from the State University of New York at Stony Brook chose to receive either payment or a research credit in a psychology course for their participation in this experiment. All participants were 18 years of age or older, and all identified themselves as native English speakers with normal hearing. None had participated in any of the previous experiments.

3.3.1.2. Materials and procedure. Participants performed the same initial Exposure phase and the same final Test phase as in Experiments 1, 2, and 3, using the same materials and procedure. In between Exposure and Test, participants performed a verbal Mixed-phoneme Unlearning task. As in Experiments 2 and 3, they heard descriptions in either the Same voice as Exposure, or a Different voice.

3.3.1.2.1. Phase II—Unlearning. The Unlearning task here was the same card-sorting task used in Experiments 2 and 3. The only difference was the phonemic information in the descriptions heard by the participants. The descriptions for the Mixed-phoneme Unlearning included equal numbers of Corrected /s/ phoneme and Corrected /ʃ/ phoneme descriptions. Specifically, all participants (whether they had heard odd /s/ or odd /ʃ/in the Exposure phase) now heard an average of one good /s/ in half of the tangrams' descriptions, and an average of one good /ʃ/ in the other half of the tangrams' descriptions (again, see Fig. 1, for examples) . Thus, each participant heard approximately 40–50 good /s/ tokens, and 40–50 good /ʃ/ tokens, during the Unlearning. Because of this balance, no adaptation effects should be generated.

The procedure for sorting the tangram cards was identical to Experiments 2 and 3: participants were randomly assigned to hear the Mixed descriptions in either the Same voice they had heard during exposure, or in a Different voice.

3.3.2. Results and discussion

3.3.2.1. Lexical decision. Any participant with lower than 75% accuracy at categorizing either the critical or filler items was replaced. Four of the 128 participants were replaced for this reason.

Table 4 summarizes the accuracy and RT data for each type of critical item. As in the previous experiments, participants were slightly more accurate at judging natural items than at judging ambiguous items ($F(1,120) = 45.68$, $p < .001$; $F2(1,19) = 11.16$, $p = .003$). In the present experiment, they were also marginally faster (17.6 ms) at judging natural items than ambiguous ones, a difference that was significant by subjects ($F1(1,120) = 4.136$, $p = .04$) but not by items ($F2(1,19) = 2.34$, $p = .14$). Despite these slight differences, accuracy for both groups was extremely high (see Table 4), with a mean accuracy of 97.0%, indicating once again that our ambiguous items were natural-sounding.

3.3.2.2. Category identification. In contrast to Experiment 3, participants in Experiment 4 heard descriptions during Unlearning that provided correct (i.e., natural) versions of *both* /s/ and /ʃ/. This was done to ensure that our results reflect pure perceptual learning (or unlearning), and are not influenced by adaptation. As in Experiment 3, a critical comparison is between those who heard these mixed (corrected) phonemes in the Same voice as exposure versus those who heard them in a Different voice.

Once again, we find an overall effect of training condition: participants who heard ambiguous /ʃ/ during lexical decision categorized more items on our continuum as /ʃ/ than participants who heard ambiguous /s/ (61.3–46.7%; $F(1,120) = 29.96$, $p < .001$). This effect marginally interacted with Unlearning voice ($F(1,120) = 2.964$, $p = .088$): although the training effect was significant in both cases, the marginal interaction reflects the fact that it was larger for participants who heard a Different voice during unlearning ($F(1,60) = 25.66$, $p < .001$) than for participants who heard the Same voice during unlearning ($F(1,60) = 7.10$, $p = .01$)

Table 4
Experiment 4, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?ssh/	
	/s/	/ʃ/	?S	?SH
% Correct	98.8 (2.7)	98.8 (2.5)	94.7 (6.7)	95.7 (4.9)
RT (in ms)	953 (128)	927 (125)	1000 (175)	915 (138)

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words. Standard deviations appear in parentheses.

(19% versus 10%, respectively). Fig. 8 shows the results for the Same voice case and Fig. 9 shows them for the Different unlearning voice.

3.3.2.2.1. Mixed phoneme, same voice condition. When we look at the Same voice condition broken down by voice (Male training–Male test versus Female training–Female test), we find that both cases show only marginal perceptual learning effects

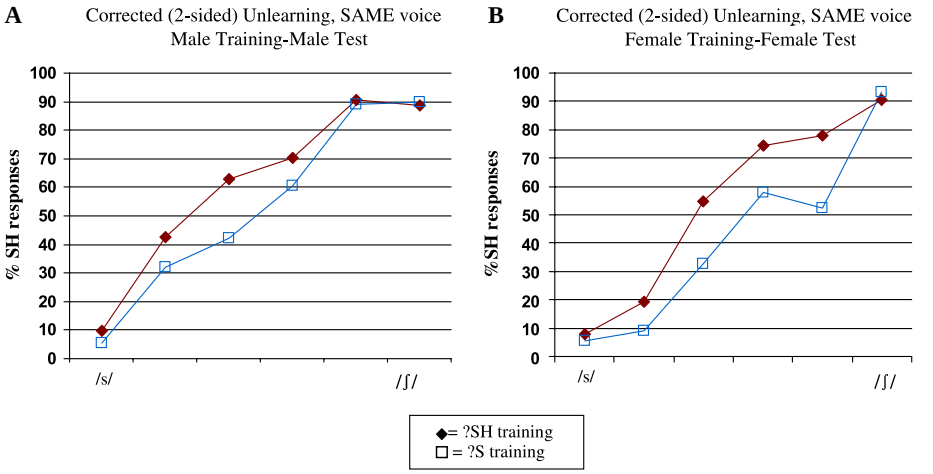

Fig. 8. Experiment 4. Mixed (corrected) phonemes during Unlearning, in the Same voice as training. The perceptual learning effect is marginal for both groups: (A) 7.4% for those who trained and tested on the Male voice and (B) 12.1% for those who trained and tested on the Female voice.

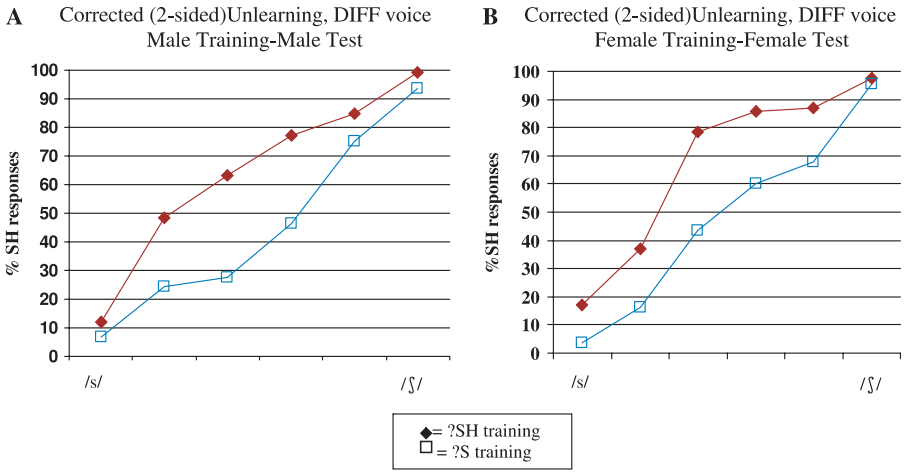

Fig. 9. Experiment 4. Mixed (corrected) phonemes during Unlearning, in a Different voice from training. The perceptual learning effect is robust for both groups: (A) 18.6% for those who trained and tested on the Male voice and (B) 19.2% for those who trained and tested on the Female voice.

(about 9%) (Male–Male $F(1,30) = 3.694$, $p = .064$; Female–Female $F(1,30) = 3.780$, $p = .061$).

3.3.2.2.2. Mixed phoneme, different voice condition. In contrast, participants who heard a Different voice during unlearning showed significant perceptual learning effects of about 19% ($F(1,30) = 11.544$, $p = .002$ and $F(1,30) = 14.332$, $p = .001$ for Male training–Male test and Female training–Female test, respectively).

These data confirm and clarify the results of Experiment 3. By removing potential adaptation influences, we find that Same voice Unlearning occurs, but not to the same extent as in Experiment 3; the difference reflects adaptation in Experiment 3. Both experiments suggest that perceptual learning adjustments may be maintained in a speaker-specific way, as a Different unlearning voice was quite ineffective. The data also indicate that once adjustments to a perceptual representation are made, they are quite stable: further input by the same speaker can attenuate the effects, but it does not fully eradicate what has been learned.

3.3.3. Preliminary Summary—Parts I and II

Across these experiments, we find an extremely stable and robust perceptual learning effect. Fig. 10 summarizes the perceptual learning effect across all of the conditions presented in Experiments 1, 2, and 4 (the results from Experiment 3 are not included because, as discussed, it is not clear to what extent they reflect adaptation as opposed to perceptual (un)learning). The figure makes it clear that perceptual learning is extremely stable—once a given pronunciation is learned for a given voice, even subsequent correcting input by that voice does not attenuate it very much. In Part III, we examine whether this learning generalizes to other voices.

4. Part III: Perceptual learning and unlearning: different speakers at exposure and test

4.1. Transfer of Perceptual Learning effects to a different speaker

One goal of the current study was to examine whether (and to what extent) the perceptual learning and unlearning results would transfer to a different speaker at test. If perceptual learning is a mechanism used to adjust to the particular properties of an individual speaker, then such effects should not transfer to test items in a very different voice. In Parts I and II, to focus on the "resetting" question, we only discussed the results that came from cases in which the training and test stimuli were presented in the same voice (Male–Male or Female–Female). In Part III, we now focus on the speaker-specificity question, and take advantage of the fact that participants in all four experiments also performed the categorization test on tokens from a different speaker than the one they had been exposed to during the lexical decision task.

In the interests of expositional clarity and brevity, we present these data in a summary form comparable to the summary provided in Fig. 10. This approach provides a profile of the perceptual learning and unlearning effects, across the six critical conditions that we tested: immediate testing (no unlearning), visual (silent) filler task,

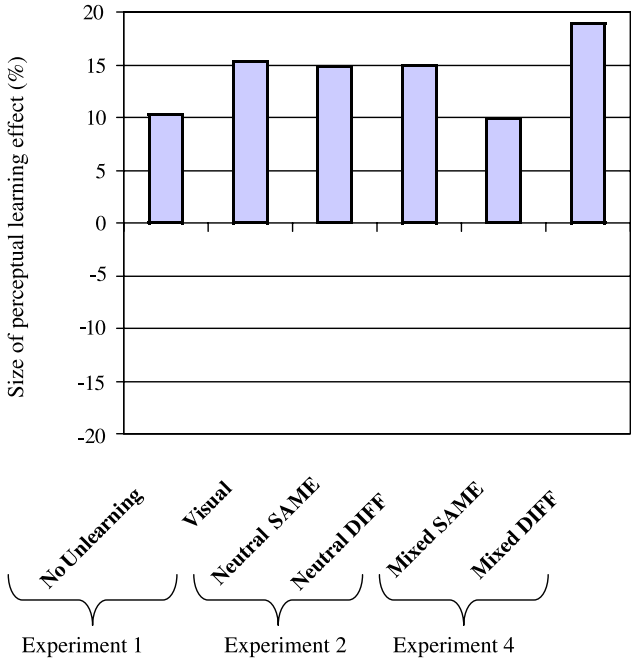

Fig. 10. Summary of the size of the within-voice perceptual learning effect for all groups from Experiments 1, 2, and 4.

unlearning without relevant phonemes (Same or Different voice than training), and unlearning with the relevant phonemes (Same or Different voice than training). Note that because the training and testing voices were always different here, unlearning in the Same voice necessarily implies that the unlearning voice is Different than the voice of the test items (and vice versa). Fig. 11A presents this summary for cases in which the training stimuli were in the Female voice and the test stimuli were in the Male voice; Fig. 11B presents the results when the roles of the two voices were reversed.

Surprisingly, the pattern of results was quite different for these two instances of mismatching voices at training and test. As shown in Fig. 11A, participants who trained on the Female voice and were then tested on the Male voice showed a pattern of results similar to the within-voice data described for Experiments 1–4 above. Although the shifts for the between-voice cases were generally smaller than those shown in Fig. 10, they are consistently in the same direction, and a few are in the same 10–15% range. An analysis of variance confirmed that across the six conditions, there was a robust perceptual learning effect, $F(1,212) = 12.234$, $p < .001$; the variations among the six conditions were not significant, $F(5,212) < 1$.

These results are quite different than those for the participants who trained on the Male voice and were tested on the Female voice. As suggested by the pattern in Fig. 11B, across the six conditions there was no perceptual learning, $F(1,212) < 1$;

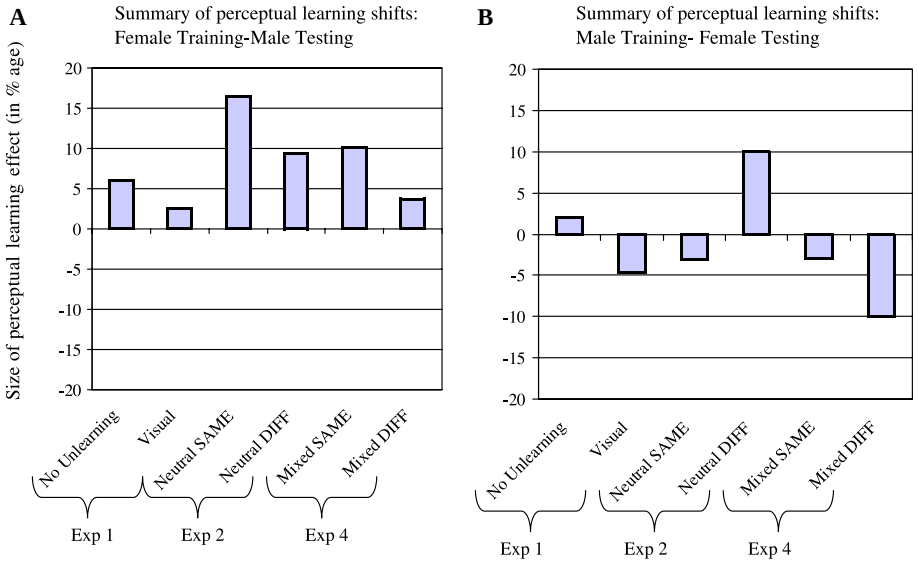

Fig. 11. Summary of the size of the cross-voice perceptual learning effect for Experiments 1, 2, and 4: (A) participants who trained on the Female voice and tested on the Male voice show perceptual learning effects, (B) those who trained on the Male voice and tested on the Female voice showed either an absence of perceptual learning effects, or effects that went in the opposite direction from perceptual learning.

there was also no significant difference among the six conditions, $F(5,212) = 1.39$, n.s. The cross-voice data, thus, seem to be in conflict with one another, with the Male training–Female test case suggesting speaker-specific adjustments, and the Female training–Male test case suggesting generalization. Why would the two combinations produce such different results?

4.2. Resolving the cross-voice asymmetry: Acoustic analyses

We have suggested elsewhere (Kraljic & Samuel, in press) that perhaps the perceptual system learns by making use of and representing the information best afforded by different types of phonemes—in other words, that abstractly defined phonemes such as stop consonants are adjusted in an abstract way, and that more acoustically defined phonemes such as fricatives are adjusted in a more specific, acoustically based way. If this is indeed the case, then perhaps unintended acoustic differences between our Female and Male training and test items led to different training-test mappings for these voices. We therefore measured the spectral mean of each fricative on the test continua (in both voices); we also obtained spectral means for the fricative in a random subset (25%) of the training items in each voice. Spectral means are a measure of one of the defining properties for fricative classification, and a main parameter for distinguishing the fricative /s/ from /ʃ/ (Jongman, Wayland, & Wong, 2000). They can be obtained by excising a portion of the relevant fricative (we used

the middle 75% of each of our critical /s/ and /ʃ/ test and training items), and using a sound editing program (see Boersma & Weenink, 2004) to obtain a single number for each item. This number represents the mean frequency of the excised portion's spectrum. The spectrum of the frication of /s/ is systematically higher than the corresponding spectrum for /ʃ/, and that for females is generally systematically higher than that for males (although individual differences across speakers' voices will also determine how high or low their fricatives' spectral means are). Fig. 12 summarizes the results of our spectral analyses.

The pattern of means supports our hypothesis that listeners use acoustic information both to guide their perceptual learning, and to decide which voices to apply those adjustments to. The average spectral mean of the Male training items is 4901 Hz for the ambiguous /s/-training items, and 4969 Hz for the ambiguous /ʃ/-training items. These means are virtually identical to each other ($t = .19$, $p = .86$) and to the mean of the Male test items (4943 Hz), $t = .02$, $p = .98$. Similarly, the average spectral mean for the Female /s/ and /ʃ/ training items are virtually identical to one another (5432 and 5383, respectively; $t = .24$, $p = .82$) On the other hand, these Female training items are significantly different in spectral mean from the Female test items (which have an average mean of 6099; $t = 3.26$, $p < .01$). The Female training items fall in between the Female and the Male test items. In fact, the frequency of the Female training items is closest to the /s/ end of the Female test continuum, and to the /sh/ end of the Male test continuum.

These measures are compatible with our perceptual learning results, in which we find that Female training transfers to the Male voice (presumably because the Female training stimuli are spectrally relatively close to the Male testing stimuli), but the Male training does not transfer to the Female voice (because Male training and test items are virtually identical in average spectral mean, and relatively distant

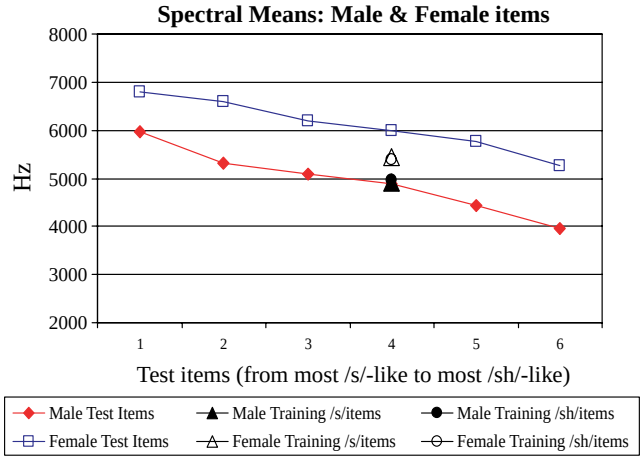

Fig. 12. Spectral means for the training and test items in both the Male and Female voices.

from the Female test stimuli). Therefore, we have evidence that the perceptual system uses acoustic cues to adjust and apply phonemic representations.[1]

5. General discussion

For the last half century, the defining issue in speech perception research has been the "lack of invariance" problem (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Hoffman, & Griffith, 1957): the acoustic realization of a given word varies substantially both across speakers and within speakers. Considerable effort has been invested in exploring ways to normalize or filter this variable signal to match it to hypothesized stable perceptual representations, such as prototypes (Samuel, 1982) or perceptual magnets (Kuhl, 1991). These approaches have been quite useful, but will probably prove to be insufficient.

The current study has examined the properties of a complementary possible solution to the problem of mismatching signals and representations. Rather than align the signal to match the representations, the representations can be adjusted to align with the input. A burgeoning literature on such perceptual learning effects for speech suggests that this approach is promising (e.g., Bertelson et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, in press; Maye et al., 2003; Norris et al., 2003; Vroomen et al., submitted).

A theoretical shift to dynamic rather than static perceptual representations brings a new set of important questions, one of which is the "resetting" problem: if a non-standard input signal prompts a changes in a listener's perceptual representation, how does the system return to its prior, presumably optimized, settings? The extensive empirical effort of the current study examined a wide set of possibilities. The simplest potential mechanism is time itself: if perceptual adjustments are "local" tags or transient rescalings, then within a short time after the triggering input is removed, the system should revert to the longer-term settings. This is the pattern that has been observed for speech identification shifts induced by selective adaptation, with such shifts evaporating within a half an hour (Harris, 1980). Our first experiment clearly demonstrated that this pattern does not obtain for lexically induced perceptual learning. Here, the shifts after a 25-min silent period were numerically larger than shifts measured immediately.

The remaining conditions of the present study tested a range of potentially resetting speech input types. These types can be characterized as the factorial crossing of Source (same versus different voice in training/test) and Relevance (absence versus presence of good tokens of the affected phoneme). Of the four

[1] Although we do not present them here, we also have measured the skewness of the training and test items (skewness refers to the spectral tilt of a frequency distribution—positive skewness suggests a concentration of lower frequency energy, and negative skewness suggests a concentration of higher frequency energy; Jongman et al., 2000). The pattern for the skewness measure converges with what we have reported for the spectral means; the Male training and test items have virtually identical skewness measures, while the Female training items actually fall somewhat closer to the Male test items than to the Female test items.

possible combinations, three clearly had no resetting effect whatsoever. As in the silent condition, perceptual effects in these cases were consistently larger than in the immediate test. Only one combination seemed at all effective: hearing "good" tokens of the critical phoneme, from the same Source that had provoked the perceptual shifts. Even in this case, the resetting was very weak. Shifts in this condition were still marginally significant, and were marginally smaller than in the three other combinations. Collectively, these results suggest that lexically driven perceptual learning causes relatively robust changes in the underlying phonemic codes.

As noted, the durability of these changes contrasts with another widely studied speech aftereffect—selective adaptation. The distinctiveness of these effects has been elegantly demonstrated in a study by Bertelson et al. (2003), and in followup work by that group (Vroomen et al., submitted; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004). In these studies the perceptual learning is driven by audiovisual manipulations, rather than lexicality, but the effects seem comparable. Acoustically ambiguous critical phonemes (midway between /b/ and /d/) are paired with a visual presentation of a face mouthing either a /b/ or a /d/. As shown by the large literature on the "McGurk effect" (McGurk & McDonald, 1976), this presentation can produce a strong auditory percept that is driven by the visual information. Bertelson et al. showed that in addition to this immediate perceptual effect, such audiovisual presentation also produced perceptual learning: subjects subsequently treated the acoustically ambiguous stimuli as members of the category they had "learned" they belonged to. This expansion of a phonemic category is entirely analogous to the expansion observed here using lexical guidance, rather than audiovisual.

Critically, this effect is in the opposite direction from selective adaptation effects. Adaptation *reduces* report of ambiguous phonemes; perceptual learning increases it. Bertelson et al. showed that when the initial exposure is to unambiguous exemplars of /b/ or /d/, the subsequent shifts in categorization followed the standard adaptation pattern: exposure to /b/ reduced report of /b/, and exposure to /d/ reduced report of /d/. In followup work (Vroomen et al., 2004; Vroomen et al., submitted), these authors have also shown differences in how the two effects build up and decay. Essentially, perceptual learning (at least in its audiovisually induced form) builds rapidly, but does not grow over large numbers of presentations; selective adaptation builds a bit more slowly, but continues to grow considerably in strength with larger numbers of adapting trials (see also Simon & Studdert-Kennedy, 1978).

These differing patterns have important implications for an ongoing debate about the architecture of the system underlying spoken word recognition. Theorists generally agree that multiple levels of representation are involved in the perception of spoken words; most commonly, features, phonemes and/or syllables, and lexical representations are hypothesized. The disagreement arises in the communication among these levels. Autonomous theories (e.g., Massaro, 1989; Norris, McQueen, & Cutler, 2000) posit strictly bottom up information flow. Interactive theories (e.g., Magnuson, McMurray, Tanenhaus, & Aslin, 2003; McClelland & Elman,

1986; Samuel, 1981), in contrast, assume that there is also top-down information flow, notably from lexical representations to phonemic codes.

In this debate, critical test cases involve situations in which lexical influences on phonemic perception are indirect, as these cases cannot be explained by even the most flexible autonomous model (Merge; Norris et al., 2000). Lexically driven compensation for coarticulation (Elman & McClelland, 1988; Magnuson et al., 2003; Samuel & Pitt, 2003) provides such evidence. A second class of such results comes from selective adaptation shifts when the adapting stimuli are acoustically neutral, but attain an effective phonemic identity through lexical context (Samuel, 1997, 2001). For example, Samuel (1997) had listeners identify stimuli from a /b/–/d/ test series, after adaptation with words like "exhi*ition" or "alpha*et", or words like "aca*emic" or "psyche*elic"; the /b/ or /d/ in each adaptor was replaced by white noise. Through lexically driven phonemic restoration, listeners perceived the missing stops, and these stops reduced report of /b/ (with adaptors like "alpha*et") or /d/ (adapted with, e.g., "aca*emic"). In Samuel's (2001) follow-up study, the adaptors were similar to the training items in the current study, with lexical context providing the interpretation of an ambiguous /s/–/ʃ/ mixture in items such as "arthriti?" or "demoli?". Again, repeated presentation of /ʃ/-based adaptors reduced report of /ʃ/, compared to adaptation with /s/-based stimuli, even though the fricative mixture was physically identical in the two cases. In both studies, only top-down constraint from the word contexts could be producing the identification of the /b/ or /d/ (Samuel, 1997) or the /s/ or /ʃ/ (Samuel, 2001), with these phonemes in turn producing the observed adaptation shifts. Critically, as Bertelson et al. (2003) have shown, adaptation shifts are in the opposite direction from those found through perceptual learning. For example, perceptual learning due to exposure to items like "arthriti?" increases report of /s/, but adaptation with such items decreases /s/ report. Nonetheless, Norris et al. (2003) argued that the perceptual learning effect could explain the adaptation results, thereby obviating the need to assume lexical influences on the perception of phonemic information (in their view, the influences are on phonemic representations, not perception). To reconcile their results with those of Samuel (2001), they said "It is also possible, however, that the same kind of perceptual learning did take place in the two studies, but, because of the selective adaptation procedure in the Samuel (2001) study, the direction of the effect was the opposite to that observed here" (p. 233). It is not clear what to make of the notion that the direction of the effect would be the opposite of what it normally is.

There is a more subtle version of the perceptual learning account that Vroomen et al. (2004) have advanced. They suggest that perhaps perceptual learning occurs rapidly, causing the neutral critical phonemes to become better exemplars of the lexically defined category; this newly learned percept can then function as a real phoneme, and produce the contrastive adaptation effect. In fact, Vroomen et al. reanalyzed the adaptation effects found by Samuel (2001), and showed that on the first block of phoneme categorization, shifts did go in the perceptual learning direction; the shifts in the remaining 23 blocks were in the opposite (adaptation-based) direction.

The adaptors in Samuel (2001) were comparable to the training stimuli used in the current study, unlike the noise-based stimuli in Samuel (1997). Thus, finding perceptual learning with them makes sense: both perceptual learning and adaptation can occur with such stimuli, and both do. But, these results are best thought of as reflecting an early dominance of the perceptual learning effect, followed by the later dominance of the adaptation effect. This differs from the view that early perceptual learning *creates* the later adaptation. There are two problems with the latter formulation. First, such a view requires the very same stimulus (an acoustically ambiguous phoneme) to simultaneously be biased in one direction (through perceptual learning), allowing it to act as an adaptor, and in the other direction (on the categorization judgments); adaptor phases alternate with categorization every half-minute. Second, if the same trial-by-trial reanalysis is done for the results of Samuel (1997), which does not have conditions conducive to perceptual learning (white noise is not like an ambiguous /b/–/d/), consistent adaptation effects are found; there is no hint of a reversal on the initial trial(s). If lexically driven adaptation needs no help from perceptual learning in this case, it would not be parsimonious to invoke it for the very similar test in Samuel (2001). Thus, these demonstrations of top-down lexical influences on phonemic perception cannot be reduced to any consequence of perceptual learning.

The Vroomen et al. reanalysis does converge with their parametric comparison of adaptation versus perceptual learning, with the perceptual learning effect developing more rapidly than adaptation. In that context, one of the results of the current study is a bit surprising: all three of our 25-min delayed testing conditions (with no relevant speech input) produced numerically larger shifts (15.2%) than the shift for the immediate test (10.5%). This difference was not statistically significant, but it is based on results from hundreds of participants, and runs counter to the idea that these effects peak quickly. Two considerations should be kept in mind. First, Vroomen et al.'s timecourse functions are based on audiovisually driven perceptual learning for stop consonants, whereas our results are based on lexically driven learning for fricatives. Second, the observation of a larger effect does not necessarily mean that the shift has actually increased. An alternative that is at least as likely is that after some delay the shifts are more *stable*, not larger. Note that the categorization test itself involves the presentation of a large number of relevant tokens, which could produce some resetting. If the system has been actively shifting its boundaries due to the training stimuli, it may be more likely to continue to adjust them (in this case, based on the categorization items) than if some time has passed since the shifting occurred. In fact, we have evidence from experiments we have done with stop consonants (Kraljic & Samuel, in press) that suggests such a drop in the shift size as more test stimuli are heard.

As noted, we had not anticipated finding larger shifts after a delay than in an immediate test. One other unexpected finding in the current study may offer some valuable theoretical leverage. In Part III, we found that training with words in the Female voice produced consistent perceptual learning when Male test tokens were used, but the reverse was not true; effects were weak or nonexistent when testing with Female tokens after Male training stimuli. The between-voice manipulation

of training versus test stimuli was designed to assess whether a main purpose of perceptual learning is to allow the listener to adapt to any idiosyncrasies of a speaker. The disparate results for the Male–Female versus the Female–Male cases indicate that no simple answer is available for this question. Moreover, the results from the two other studies to use this manipulation also yielded differing outcomes. Eisner and McQueen (2005) argued for the speaker-adjustment position, based on having found no transfer when Female training stimuli preceded Male test tokens (using a contrast of /f/ versus /s/). In contrast, in our study of perceptual learning of stops (/d/ versus /t/), we observed full transfer between Male and Female voices.

The spectral analyses of the current study offer a possible resolution of all of these apparently conflicting findings. If perceptual learning is based on *acoustic* properties, all of the observed results may actually be consistent. We have already explained how the asymmetry found in Part III can be accounted for from this perspective. For the /d/–/t/ contrast (Kraljic & Samuel, in press), the acoustic cues distinguishing /d/ from /t/ are primarily temporal (e.g., duration of pre-release silence, duration of aspiration). As such, these will be shared across Male and Female voices, leading to the observed speaker generality. Because the /f/–/s/ distinction used by Eisner and McQueen (2005) is primarily spectral, it would be expected to differ between their Male and Female voices, producing speaker specific effects. In fact, when Eisner and McQueen spliced the vowel from their Male test tokens onto the fricatives from the Female voice, they found reliable perceptual learning effects after Female training stimuli. Note that such test stimuli sound unambiguously like the Male voice, but the acoustics of the consonants match the acoustics of the training tokens.

Thus, an acoustically grounded mechanism for perceptual learning can account for the various patterns of speaker-specific and speaker-general effects. Of course, in the real world, rather than in contrived laboratory settings that include strange splicings and mixtures, acoustic cues will generally covary with speaker identity. As a result, a system that uses variations in the acoustic patterns will in effect also be adapting to the idiosyncrasies of the speaker.

As we discussed in Section 1, models of speech perception have primarily focused on how invariant linguistic representations might be accessed despite all the variations inherent in speech. The present results suggest that it is not the processes that must adapt to account for such variability, but the representations themselves. We argue that phonemic representations are dynamic and flexible, and incorporate specific information about the speech, and the speaker, that a listener is exposed to. This finding ties to the more general literature on perceptual memory for voices (Goldinger, Pisoni, & Logan, 1991; Mullinex & Pisoni, 1990; Nygaard & Pisoni, 1998), which finds that memory for speakers appears to aid perception and interpretation of the speech signal, and argues that speech perception and word recognition models need to incorporate partner specific voice information (see Goldinger, 1998, for a model which does this). This body of work provides interesting evidence that speakers are important to listeners, and it supports the idea that information about partners may well be a basic part of

ordinary memory representations, and not incorporated at some later extralinguistic processing stage.

More generally, the present studies have implications about language processing in dialogue. With the exception of the work on cross-talker variability discussed in the preceding paragraph, speech research has virtually ignored any role for remembering a conversational partner's voice in language processing. On the other hand, there is a generous body of research at the syntactic, lexical, and other higher levels devoted to proposing how language is impacted by experience with a conversational partner. Although this research has predominantly focused on linguistic *behaviors* (e.g., showing that a speaker will formulate the same message differently for different listeners), there have been some attempts to address the processes that give rise to those behaviors. These attempts have resulted in two very general lines of theory. One of these argues that speakers and addressees create and maintain 'models' of one another, which inform and constrain linguistic processing (see in particular Clark & Carlson, 1982; Clark & Marshall, 1978). The other theory argues that elaborate partner models are unnecessary for most cases of language use, and that linguistic adaptations result from an automatic priming mechanism that causes output to be based on the most recent input (e.g., Brown & Dell, 1987; Garrod & Anderson, 1987); thus, a person's speech may appear to be adjusted to reflect partner knowledge, but it really simply reflects the most recent speech encountered.

The present data suggest that partner adaptations, at least at the perceptual level, are due neither to elaborate models nor to an automatic priming mechanism, but that instead experience with a partner influences subsequent linguistic choices via long-term changes to the representations themselves. The representational adaptations we have shown in these studies persist even with intervening, correcting input, which lends further support to the idea that adaptations to a partner cannot simply reflect local coordinations (e.g., priming), but that instead they reflect information about partners that have become a basic part of ordinary memory representations. We suggest that such adaptations occur on the basis of a defining feature of the representation (in this case, the acoustics) and then are recalled when the acoustics at input match (or are similar to) the newly expanded category.

Acknowledgments

This material is based upon work supported by a National Science Foundation Graduate Fellowship and by NSF Grant No. 0325188 and NIH Grant R0151663. We are extremely grateful to Donna Kat for all of her ideas and feedback, and for her invaluable technical support, and to Glenn Spaeth and Alex Cristodoulou for their help with running many, many subjects. We thank Steve Goldinger, Dennis Norris, and an anonymous reviewer for their helpful suggestions.

Appendix A. Critical and filler words (60 nonwords are in parentheses next to the word they were created from; the additional 40 nonwords are listed separately)

Critical /s/ words (S)
Arkansas
eraser
coliseum
compensate
democracy
dinosaur
embassy
peninsula
Episode
hallucinate
legacy
Literacy
medicine
obscene
personal
Parasite
pregnancy
reconcile
rehearsal
Tennessee

Critical /ʃ/ words (SH)
ambition
beneficial
brochure
commercial
crucial
efficient
flourishing
glacier
graduation
impatient
initial
machinery
negotiate
official
parachute
pediatrician
publisher
reassure
refreshing
vacation

Filler words (nonwords)
Accordion (igoldion)
America (anolipa)
Annoying (imoyem)
Armadillo (alnadiro)
Bakery (pakelo)
Ballerina (galliwinou)
Blueberry (pluepelai)
Bullying (pourilar)
Camera (ganla)
Crocodile (klogodar)
Darken (perkum)
Directory (tilegkalo)
Document (pogunemd)
Domineering (konimeelum)
Dynamite (tymolipe)
Embody (enpaiki)
Gardenia (kaldemia)
Grammatical (kloumidiger)
Gullible (kuradel)
Hamburger (hintarber)
Honeymoon (hominaim)
Hurdle (hilder)
Identical (itempider)
Ignite (aknid)
Immoral (irimel)
Inhabit (emhoutic)
Knowingly (mowery)
Laminate (wonimtic)
Legally (weekary)
Liability (riakirity)
Lobbying (woppakin)
Lunatic (rumatik)
Lyrical (ryligal)
Manually (namuery)
Marina (nawinow)
Melancholy (neramgory)
Membrane (nempring)

Memory (nomeray)
Metrical (nekridal)
Military (niritaly)
Momentary (nomemtoly)
Napkin (mibgem)
Negate (mikid)
Outnumber (admunker)
Panicky (bimikay)
Parable (baliber)
Parakeet (bawaseet)
Pineapple (bimobel)
Platonic (kradomet)
Remedial (lenediaw)
Romantic (wonontic)
Tactical (dadigal)
Titanium (bikanian)
Turbulent (durkuwomt)
Tutorial (datiliar)
Umbrella (omplero)
Warrantee (rawamtee)
Wealthy (lirthy)
Withdrawal (rikmaral)
Wrinkle (lindel)

Additional Nonwords
Acominig
Aigi
Ailounam
Amalar
Anemer
Bamtel
Bliparg
Gairelom
Galliwinou
Gerbualo
Geypalg
Gilday
Gondimually
Gonedial
Halomimoc
Hiliun
Ibirak
Imdalier
Ithomel
Kegimel
kelabidel
kermimer
kerkrun
lilgrai
logelai
loubel
maidnow
marody
omperog
pirugalo
rakil
rengimer
rimkuwar
tamical
tounamlemp
umikory
ungelnin
waiper
wojalto
youmgel

Appendix B. Critical words and the /?sʃ/ mixtures used for each word

Word	Mixture	
	Male	Female
S-items		
Arkansas	60	50
Coliseum	50	50
Compensate	50	40
Democracy	50	40

Appendix B (*continued*)

Word	Mixture	
	Male	Female
dinosaur	40	40
eraser	60	50
embassy	50	40
episode	70	50
hallucinate	50	50
legacy	50	30
literacy	70	60
medicine	40	50
obscene	50	50
personal	60	50
parasite	50	50
pregnancy	60	60
peninsula	40	50
reconcile	40	50
rehearsal	50	50
Tennesee	50	60
Average mixture chosen:	52	48.5
SH-items		
ambition	60	30
beneficial	40	50
commercial	50	50
brochure	80	40
crucial	70	20
efficient	20	30
flourishing	60	50
glacier	70	50
initial	50	30
machinery	40	35
negotiate	60	40
official	50	35
parachute	50	35
pediatrician	40	45
publisher	60	50
reassure	50	55
refreshing	70	55
vacation	70	45
graduation	50	50
impatient	20	30
Average mixture chosen:	53	41.25

Note that the number refers to the percentage of /ʃ/ in the mixture (out of 100).

References

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. (2004). Praat: doing phonetics by computer (Version 4.2.07) [Computer program]. Retrieved June 24, 2004, from http://www.praat.org/.

Bradlow, A. R., Akehane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985.

Bradlow, A. R., & Bent, T. (2003) Listener adaptation to foreign accented English. In M. J. Sole, D. Recasens, & J. Romero (Eds.), *Proceedings of the XVth international congress of phonetic sciences* (pp. 2881 2884). Barcelona, Spain.

Bradlow, A. R., Pisoni, D. B., Akehane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*(4), 2299–2310.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition, 75*, B13–B25.

Brennan, S. E. (1999). The vocabulary problem in spoken dialogue systems. In S. Luperfoy (Ed.), *Automated spoken dialogue systems*. Cambridge, MA: MIT Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1482–1493.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology, 19*, 441–472.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2005). Real-time reference resolution in a referential communication task. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Processing world-situated language: Bridging the language as product and language as action traditions*. Cambridge, MA: MIT Press.

Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language, 58*, 332–373.

Clark, H. H., & Marshall, C. R. (1978). Reference diaries. In D. L. Waltz (Ed.). *Theoretical issues in natural language processing* (Vol. II, pp. 57–63). New York: Association for Computing Machinery.

Clarke, C. M. (2000). Perceptual adjustments to foreign-accented English. In Indiana University's Research on Spoken Language Processing Progress Report, No. 24.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Society of Language, 46*, 49–70.

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

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

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143–165.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181–218.

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*, 251–279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 152–162.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*(3), 421–433.

Harris, L.B. (1980). An assessment of current hypotheses concerning selective adaptation using phonetic stimuli. Unpublished Masters thesis, SUNY Binghamton.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*(3), 1252–1263.

Kolers, P. A. (1976). Pattern-analyzing memory. *Science, 191*, 1280–1281.

Kraljic, T., & Samuel, A. G. (in press). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*.

Kucera, F., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93–107.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*, 78–106.

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

Liberman, A. M., 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*, 538–568.

Magnuson, J., McMurray, B., Tanenhaus, M., & Aslin, R. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27*, 285–298.

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

Maye, J., Aslin, R., & Tanenhaus, M. (2003). In search of the weckud wetch: Online adaptation to speaker accent. In *Proceedings of the 16th Annual CUNY Conference on Human Sentence Processing,* March 27–29, Cambridge, MA.

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

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Cristophe, A., & Pallier, C. (1993). Understanding compressed sentences: The role of rhythm and meaning. *Annals of the New York Academy of Sciences, 682*, 272–282.

Mullinex, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

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

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*(1), 42–46.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474–494.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31*, 307–314.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology, 18*, 452–499.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*, 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12*, 348–351.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48*, 416–434.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 915–930.

Simon, H. J., & Studdert-Kennedy, M. (1978). Selective anchoring and adaptation of phonetic and nonphonetic continua. *Journal of the Acoustical Society of America, 64*, 1338–1357.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*, 339–363.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (submitted). The build up of recalibration and selective adaptation in auditory-visual speech perception.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication, 44*(1–4), 55–61.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates.