Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 56 (2007) 1–15
Journal of Memory and Language
www.elsevier.com/locate/jml

Perceptual adjustments to multiple speakers

Tanya Kraljic [a,*], Arthur G. Samuel [b]

[a] University of California at San Diego, Center for Research in Language (CRL), Cognitive Science Department, 9500 Gilman Drive, La Jolla, CA 92093-0526, USA
[b] Department of Psychology, State University of New York at Stony Brook, Stony Brook, NY 11794-2500, USA

Received 15 October 2005; revision received 27 June 2006
Available online 18 September 2006

Abstract

Different speakers may pronounce the same sounds very differently, yet listeners have little difficulty perceiving speech accurately. Recent research suggests that listeners adjust their preexisting phonemic categories to accommodate speakers' pronunciations (*perceptual learning*). In some cases, these adjustments appear to reflect general changes to phonemic categories, rather than speaker-specific adjustments. But what happens when listeners encounter multiple speakers with different pronunciations? We exposed listeners to two speakers who varied in their pronunciation of a particular phoneme (Experiment 1: /d/ or /t/; Experiment 2: /s/ or /ʃ/). Listeners then categorized sounds on /d/-/t/ or /s/-/ʃ/ continua, in the same two voices. The results suggest that perceptual experience leads to very different learning for different types of phonemic contrasts. For fricatives, perceptual learning was speaker-specific: The system was able to maintain multiple different representations simultaneously. In contrast, perceptual learning for stop consonants resulted in more general changes that required the system to re-adjust when a new pronunciation was encountered.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Perceptual learning; Speech perception; Partner effects; Adjustments; Multiple speakers; Priming

Anyone who has used a speech-to-text system knows that most systems are not equipped to handle multiple speakers: to get even moderately successful performance, the system has to be trained on the user's voice for an hour or two. Thereafter it will be able to translate that voice's speech to text, but not the speech of other people.

Humans, on the other hand, are very good at adjusting to multiple speakers. At any given gathering we might be involved in a conversation with several people who have different foreign accents or who speak different American dialects. Consider a family party with loud conversation around a table with native Croatian, German, and Spanish speakers, people with Long Island dialects, and a cousin from Boston: Imagine a speech recognizer trying to keep up with that, even at a slow pace! And yet in conversation we rarely experience comprehension difficulty when one speaker switches rapidly to another. How are our perceptual systems able to handle such speaker-driven variability so effectively?

Recent research on perceptual learning offers one potential solution: It appears that listeners adjust their phonemic representations to reflect the speech that they are exposed to. In their seminal study, Norris, McQueen, and Cutler (2003) exposed Dutch listeners to an ambiguous sound midway between an /f/ and an /s/; this

* Corresponding author. Fax: +1 858 822 5097.
E-mail address: tkraljic@crl.ucsd.edu (T. Kraljic).

0749-596X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2006.07.010

sound occurred at the end of Dutch words that normally end in /f/ (e.g., *witlof*, which means chicory), at the end of words that normally end in /s/ (e.g., *naaldbos*, pine forest), or at the end of nonwords. Listeners who heard the ambiguous sound in the context of /f/-final words later categorized more items on an /εf/-/εs/ continuum as /f/, whereas listeners who heard this sound in /s/-final words categorized more items on the same /εf/-/εs/ continuum as /s/, suggesting that perceptual learning results in phonemic categories that are expanded to match the input. Hearing the ambiguous sound in nonwords produced no such shift. Thus, listeners appear to use lexical knowledge to dynamically tune their phonemic representations according to the speech and pronunciations they hear.

Other work has replicated and extended these findings to various types of synthetically created speech (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Maye, Aslin, & Tanenhaus, 2003). Maye, et al., for example, had participants listen to a story read either in Standard English, or in an 'Accented English' that they had created. In the Accented English version, some words contained lowered front vowels (e.g., the speaker would pronounce *wetch* instead of *witch*). Listeners who had heard this Accented English were subsequently more accurate and faster at responding to words with lowered front vowels than were listeners who had heard the same stories read in Standard English. Further, whatever perceptual change occurred after exposure to accented words did not result in more broad or sloppy phonemic categories that simply allowed for more noise in pronunciation; rather, the change was specific to the direction of the accent.

We also now know that perceptual learning is not restricted to cases in which the *lexical* context guides interpretation of 'odd' or different pronunciations: Visual context can also serve to constrain the interpretation of phonemes, and therefore results in perceptual learning that is comparable to Norris, et al.'s original finding (e.g., Bertelson, Vroomen, & de Gelder, 2003).

Norris et al. (2003) suggested that the function of perceptual learning was to adapt to a particular speaker's accent. But what happens when we encounter multiple speakers with different pronunciations of the same sound, as might happen during a dinner party? Do we maintain separate phonemic representations for each person's pronunciation that we quickly access when that person speaks? Or do we continually adjust the same phonemic representation to reflect each person's pronunciation for a short time, only to be re-adjusted with the next person's pronunciation?

Some recent findings suggest that the answer may be quite different depending on the sound contrast that is being varied. In two projects in our laboratory (Kraljic & Samuel, 2005, 2006), listeners were exposed to speakers whose pronunciation of a particular sound (for example, either a /d/ or a /t/) was ambiguous (i.e., it was midway between /d/ and /t/). Following the Norris et al. (2003) paradigm, the lexical context in which the sound occurred determined how listeners would come to perceive it; for example, some listeners heard the ambiguous sound in place of /d/ (in words such as *crocodile*), and others heard the same sound in place of /t/ (in words such as *cafeteria*).

After this exposure, all participants were tested for perceptual learning on two vowel-consonant-vowel continua, in which the consonant ranged from very /d/-like to very /t/-like; for each item of the continua, listeners indicated whether the sound they heard was /d/ or /t/. The presence and extent of perceptual learning was assessed by differences in listeners' categorizations as a function of whether they had heard the ambiguous fricative in /d/ or in /t/ lexical context. Those listeners who had been exposed to the ambiguous sound in /d/ words (e.g., *crocodile*) perceived more items on the continua as /d/; those who had been exposed to the same sound in /t/ words (e.g., *cafeteria*) perceived fewer items on the continua as /d/.

Critically, one of the continua was presented in the Same voice the listener had heard during exposure; the other continuum was presented in a Different voice. If perceptual learning results in speaker-specific adjustments (equivalent to learning that *This is Speaker X's /d/*), then we should not see perceptual learning for continua presented in a Different voice than the one listeners had heard during exposure. If, on the other hand, the adjustments reflect some more general tuning of the phonemic representation that is not tied to the speaker (e.g., *Accept a wider range of voice onset times as voiced stops*), we should see perceptual learning for a Different voice to the same extent that it occurs for the Same voice.

The results were clear: When the critical items contained manipulated stop consonants (/d/ and /t/), as just described, there was significant perceptual learning for both the Same voice that listeners had heard during exposure and for a Different voice heard only during the categorization test. In fact, we saw significant perceptual learning even for different (and previously unheard) phonemes that shared the same timing feature as the critical sounds (i.e., perceptual learning generalized to /b/ and /p/ as well; see Kraljic & Samuel, 2006).

The most likely explanation for such speaker- and phoneme- generalization is that learning occurs for the contrast being manipulated, and the information that particular contrast affords will determine the scope of the learning. For the stop consonants /d/ and /t/, the primary contrast is a temporal-voicing one (i.e., the duration of stop closure + the duration of voice-onset time, or VOT). Critically, these acoustic cues do *not* provide local, acoustic information about who is speaking

(or even what the particular phoneme is): The stop closure is simply a period of silence, while VOT (the period between the closure release and subsequent voicing onset) may have some acoustic energy, but this energy does not reliably signal a particular speaker.

The idea that listeners rely on local acoustic cues to make perceptual learning adjustments was strengthened by a second finding: When we manipulated sounds along a dimension that *should* provide local speaker-specific information, we found quite a different pattern of results than we had for the stop consonants. In Kraljic and Samuel (2005), the manipulated sounds were the fricatives /s/ and /ʃ/. The primary contrast between /s/ and /ʃ/ is a spectral-place contrast rather than a temporal-voicing one (/s/ is an alveolar sound and has a higher spectral frequency than /ʃ/, which is produced with the tongue farther back in the mouth). Critically, such spectral differences not only distinguish one phoneme from another, but also frequently distinguish one speaker from another (e.g., a male speaker from a female one). When we tested perceptual learning for this fricative contrast, we found speaker-specific perceptual learning in one of our conditions (in which listeners were exposed to a Male voice and then tested on a Female voice). In this case, listeners appeared to be learning something about a particular speaker's pronunciation of a sound (*This is Speaker X's /s/*), rather than something more general that would then be applied to a new speaker. In addition, learning was more robust, and persisted even after listeners were exposed to a 'corrected' version of the previously ambiguous fricative (Kraljic & Samuel, 2005).

This type of speaker-specific adjustment is consistent with what other researchers (Eisner & McQueen, 2005) have found when using fricatives (/s/ and /f/) as the critical phonemes. In their study, participants were exposed to a female speaker whose pronunciation of either /f/ or /s/ was ambiguous between the two phonemes; after this exposure, each participant identified items on an /ɛf/-/ɛs/ continuum. Critically, the vowel portion of the continuum (/ɛ/) was either produced by the same speaker that participants had been exposed to, or it was replaced by a different (but similar) female voice (Experiment 1) or by a male speaker's voice (Experiment 2). The fricative portion was always produced by the original female speaker. Although participants in the latter two groups believed that the continuum items were produced by a new speaker, the perceptual learning effect was obtained for all three groups. In contrast, when the entire continuum (including the fricative portion) was produced by the male speaker, no perceptual learning effect was obtained. These results also provide support for the idea that the perceptual system adapts to variation in a local speaker-specific way: i.e., that adaptations in the perceptual system seem to be driven by acoustic cues, rather than by higher-level information about a speaker's identity.

One final piece of evidence that perceptual learning relies on acoustic information comes from an unexpected finding in Kraljic and Samuel (2005), where (as we discussed) the critical phonemes (/s/ and /ʃ/) varied along a spectral dimension. As noted above, we found that when listeners were exposed to a Male voice and then tested on a Female voice, there was no perceptual learning, indicating that the experience had led to a speaker-specific adjustment. However, when listeners were exposed to the Female voice and then tested on the Male voice, we did find perceptual learning: As with the stop consonants, listeners were generalizing what they had learned to a new voice. Acoustic analyses of the /s/ and /ʃ/ phonemes at training and at test showed that the Female training stimuli were spectrally relatively close to the Male testing stimuli (in fact, the Female training stimuli were intermediate between the Male and the Female testing stimuli). In contrast, the Male training and test items were virtually identical in average spectral mean, and substantially lower than the Female test stimuli.[1] This would explain why the learning on the Female voice transferred to the Male voice at test, but not vice versa: The acoustic cues provided in the Female voice's critical phonemes were actually ambiguous with respect to speaker, and as a result the cues did not enable the perceptual system to differentiate between the same (female) or new (male) phonemes at test. Therefore, the perceptual system applied what it had learned to both voices.

Taken together, the previous perceptual learning work suggests that the perceptual system relies on acoustic cues to adjust phonemic representations, even if the goal is to adjust to more abstract things like individuals. Specifically, when the to-be-learned phoneme highlights a temporal-voicing contrast that does not provide local, acoustic cues to speaker, as in our stop manipulations, learning will be speaker-independent. But when it highlights a spectral-place contrast that does acoustically distinguish one speaker from another, as in one of our fricative manipulations, learning is speaker-specific.

A potential problem for a system with these properties is that relying on such local cues may be less effective when listeners have to adjust to multiple speakers at the same time (as in our initial dinner-party example). The purpose of the present experiments, then, is to explore the implications that the previous perceptual learning data have for our original question: How do listeners handle the variation present when different speakers

[1] Note that the differences in the Female training and test stimuli were a consequence of creating the two from different utterances; it seems that producing a higher vowel (/i/) immediately following the fricative (as in the test stimuli) caused the female speaker to produce the fricative with a higher frequency as well. No such tendency was found for the male speaker.

pronounce the same sound differently? There are at least two possibilities: First, listeners might maintain separate phonemic representations for each person's pronunciation. This could be true even though we have seen that the system can, and sometimes does, generalize what it has learned for one person to another (e.g., in the temporally-varied stops). Perhaps when the system is faced with adjusting to multiple pronunciations brought on by multiple speakers, it comes to rely less on purely acoustic cues and more on higher-level cues to speaker identity.

Alternatively, listeners might continually adjust the same phonemic representation to reflect each person's pronunciation, only to be re-adjusted with the next person's pronunciation. This latter possibility reflects a potential alternative explanation for the finding that perception changes in response to experience: These effects could reflect *priming*. An account that has been put forth primarily by discourse researchers to explain partner effects in conversation (e.g., Dell & Brown, 1991; Pickering & Garrod, 2004), priming is a mechanism in which apparent adjustments to particular speakers and listeners do *not* mean that such information has been represented by the linguistic system. Instead, according to this view, those linguistic forms and concepts that have recently been used (that is, heard or spoken) simply enjoy a temporary increase in their level of activation (priming). On this view, experience results in temporary *behavioral* adaptations, but not in any long term adaptations to the underlying representations themselves. The implication is that any observed adaptation to a speaker is a by-product of a system that is designed to use forms and concepts that have recently been accessed, thereby increasing communicative efficiency.

Priming would not be a sufficient explanation if the perceptual system could be shown to simultaneously adapt to multiple speakers in a speaker-specific way. There are both logical and empirical reasons for supposing that such adaptations can and do, in fact, occur. On the logical side, if the purpose of perceptual learning is to make communication more efficient, then it would make sense for the system to maintain dynamic representations that adjust to different speakers. Given that we continuously encounter speech from different people, sometimes many times within a single conversation, what we have learned for one speaker can't 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.

Empirically, in addition to the findings using fricatives, which resulted in speaker-specific learning (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), evidence at other levels of language processing suggests that listeners interpret the same referring expressions differently for different speakers, based on their experience with a particular speaker, and that this experience exerts an influence extremely early in processing (see Metzing & Brennan, 2003; Nadig & Sedivy, 2002). These findings suggest that the linguistic system does keep track of what information is relevant to our interactions with different conversational partners.

These analyses suggest a critical test: What happens when we encounter multiple speakers with opposing pronunciations of the same sound? There are three possibilities: (1) Listeners might maintain *distinct*, speaker-specific phonemic representations. If this were the case, we would expect to see significant, but opposite, perceptual learning effects for each speaker whom listeners were exposed to. (2) Listeners might retune the *same* phonemic representation each time they encounter a particular pronunciation of a phoneme, resulting in no perceptual learning effect (because the two opposing effects would cancel each other out). (3) The most recent pronunciation might override any previous retuning, resulting in a perceptual learning effect that is the same for both speakers, and that is consistent with the most recent pronunciation heard (regardless of speaker). This final possibility is what would be predicted by a priming account of perceptual learning.

Given the findings reviewed previously, it appears that perceptual experience may lead to very different learning for different types of phonemic contrasts, resulting in different outcomes that depend on the nature of the representation itself and on the nature of the information provided locally in the acoustic signal. That is, for stop consonant voicing, for which perceptual learning seems to result in general phonemic adjustments, we may find that hearing multiple pronunciations will result in a net absence of perceptual learning, because the same representation has been tuned in two directions. On the other hand, perceptual learning for spectral contrasts such as those in fricatives, which can be speaker-specific, may result in the maintenance of distinct adjustments that are applied to the appropriate speaker.

The present experiments test these hypotheses. Listeners were exposed to two speakers (Male and Female) in the context of a lexical decision task. In Experiment 1, the critical sound was a stop consonant midway between /d/ and /t/ (?dt). This sound replaced the /d/ (in words such as *croco?ile*; the ?D condition) for one speaker; a similarly ambiguous sound replaced /t/ (in words such as *café?eria*; the ?T condition) for the other speaker. Thus, for successful perception to occur, listeners must learn to perceive ?dt as /d/ when hearing one voice, but as /t/ when hearing the other voice. After this exposure, participants categorized items on an /IdI/-/ItI/ continuum in both voices. Experiment 2 was identical, except that the critical sounds were the fricatives /s/ and /ʃ/.

If listeners maintain speaker-specific representations, their categorization of the continuum items should be different for the two voices they have trained on. If, however, perceptual learning is applied more speaker-generally, it will be interesting to see the pattern of perceptual learning. Will the representations reflect an average of the two pronunciations that have been heard, or will they reflect (as a priming account would suggest) the most recent pronunciation?

Experiment 1

Method

Participants

One hundred and twelve undergraduate psychology students from the State University of New York at Stony Brook participated for a research credit or for payment. All participants were 18 years of age or older, and all were native English speakers with normal hearing.

Design

Participants were randomly assigned to one of four control groups, or to one of two experimental groups. In each case, participants performed a lexical decision task, followed by a categorization task. Control and experimental groups differed in the number of voices they heard during exposure (and therefore in the number of total items): Participants in the control groups heard a single voice during exposure (either Male or Female) and a single 'mispronounced' (i.e., ambiguous) phoneme (either ?D or ?T) that occurred in 10 critical words. Crossing the voice at exposure with the ambiguous phoneme thus resulted in the four control groups (Male?D, Female?D, Male?T, Female?T). Sixteen participants were randomly assigned to each of these four groups, for a total of 64 control participants. All of the previous studies using Norris et al.'s (2003) paradigm (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006) have used 20 critical items during the lexical decision exposure phase. Therefore, the purpose of the control conditions here was to ensure that exposing listeners to only 10 critical items (and therefore only 10 ambiguous pronunciations) is sufficient to induce perceptual learning of that pronunciation.

The remaining 48 participants were assigned to one of the two experimental groups. Participants in the experimental groups were exposed to *both* the Male *and* the Female voice during the lexical decision task, and to *both* mispronounced /d/s *and* mispronounced /t/s. Each participant in the experimental group heard 10 critical items in the male voice, and 10 critical items in the female voice. The order of presentation of the items was blocked by voice. Specifically, each experimental participant heard one block in which the /d/ was ambiguous (?D) and one block in which the /t/ was ambiguous (?T). Thus, each experimental participant either heard Male ?D *and* Female ?T or the alternative combination (Male ?T and Female ?D). Order of voice (Male first or Female first) was included as a factor for counterbalancing purposes.

Although the experimental participants were thus exposed to twice the number of items overall as control participants, the items potentially should generate perceptual learning in opposite directions for each voice. If the perceptual system is not able to make changes in a voice-specific manner, the net result for the experimental groups would be *no* perceptual learning effect.

In the Categorization Test phase, *all* participants categorized phonemes on two /IdI/-/ItI/ continua, one in the Male voice and one in the Female voice. Continua were blocked by voice, and the Order of the voices was included as a factor for counterbalancing purposes.

Materials and procedure

Phase 1—Exposure (lexical decision). Two experimental lists were created for use in the auditory lexical decision task, each with 100 words and 100 nonwords. The lists were identical except for 40 critical words.

Stimulus selection. The 40 critical words ranged in length from two to five syllables (see Kraljic & Samuel, 2006, for a complete list of words). Twenty of the words contained no /t/, and had a single instance of the critical phoneme /d/. The other twenty critical words contained no /d/; these each had a single instance of the phoneme /t/. Each control subject heard half of the critical items (10 /d/ and 10 /t/), with the choice of items counterbalanced across subjects; each experimental subject heard all 40 of the critical items, half in a Male voice and half in a Female voice (this manipulation is explained more fully below).

The two sets of critical words were matched in mean syllable length and in frequency of occurrence (Zeno, Ivens, Millard, & Duvvuri, 1995). We also selected 100 filler words that had no occurrences of /d/ or /t/. As in Kraljic and Samuel (2006), the fillers were matched to the critical words in term of stress pattern, number of syllables, and word frequency; 60 of these words were used in the lexical decision task. Finally, 100 filler nonwords were created. Each experimental participant thus heard 100 words and 100 nonwords: In the ?D block, there were 10 /d/-words (containing ?dt), 10 intact /t/-words, 30 filler words, and 50 filler nonwords; for the ?T block, there were 10 /t/-words (containing the ?dt), 10 intact /d/-words, 30 filler words, and 50 filler nonwords.

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. The recording procedure, as well as the procedure for creating the ambiguous (?dt) mixture, is described in Kraljic and Samuel (2006). Each mixture varied three cues: the relative amplitude weighting of /t/ and /d/ in the mixture (100% /t/ plus 0% /d/, down to 0% /t/ plus 100% /d/, in 5% increments), the length of ?dt after the voicing burst onset (longer aspiration favors /t/; range = 100% of the original /t/ length, down to 0% of the original /t/, in 5% increments), and the length of silence before the burst onset (longer favors /t/; range = 40–0 ms, in 2 ms steps). Using this procedure, we created 21 mixtures of /d/ + /t/ for each critical word; ultimately, we chose one of these mixtures for each item (the most ambiguous one, as rated by the authors and an independent rater) for use in the experiment.

We calculated the total duration of ?dt for the critical items for each speaker. We measured the time from the end of the vowel immediately preceding each critical /d/ and /t/ to the end of the aspiration. Note that this measure takes into account both of our timing manipulations: silence before burst onset, and burst onset to subsequent vowel. There was no difference in the average length of ?dt across the male and female voices (66.5 vs. 69.7 ms, respectively, $t(1,38) = .7$, $p = .49$).[2] We also calculated two spectral measures for each critical ?dt to ensure that the burst energy for the stops did not differ systematically across voices: We measured the spectral mean (frequency), and the Root Mean Square (RMS) amplitude from the burst onset to the onset of the subsequent vowel for each item using Praat. As expected, there were no differences between the male and the female voices in either frequency of the burst energy (4047 vs. 4603 Hz, respectively, $t(1,38) = 1.75$, $p = .08$) or in RMS amplitude of burst energy (.099 Pascals vs. .091 Pascals, respectively, $t(1,38) = .28$, $p = .77$).

As explained above, participants in the control groups heard only one of these voices, and therefore only half of the items: they heard 10 critical words with /d/, 10 critical words with /t/, 30 filler words, and 50 filler nonwords. For half of the control participants, the 10 /d/'s were replaced with the ambiguous ?dt version of the item(?D); for the other half, the /d/-words remained unchanged, but the 10 /t/'s were replaced with the ambiguous ?dt mixture (?T). These two conditions were crossed with voice of presentation (Male or Female).

Participants in the experimental groups, however, heard *both* voices and *both* ambiguous phonemes, and therefore all 200 items (i.e., 24 of the experimental participants heard Male?D + Female?T, and the other 24 heard the alternative combination of Male?T + Female?D). In order to maximize any effect of each voice's pronunciation, the voices were blocked (with the order counterbalanced). An important advantage of blocking voices is that it allows control over which Voice-Pronunciation combination each participant had been exposed to immediately before the categorization test. This enables us to test whether more recently heard pronunciations 'override' previously heard ones (as predicted by a priming account), or whether the two simply cancel one another out regardless of order.

Phase II -Category Identification. In the second phase of the experiment, all participants heard six items on a vowel-Consonant-vowel (vCv) continuum, presented in two voices. The two endpoints of the continuum (/IdI/ and /ItI/) were recorded by the same male and female speakers who produced the lexical decision stimuli. Twenty-one mixtures of each continuum were created and raters chose six consecutive tokens for each continuum. These stimuli ranged from relatively /d/-like to relatively /t/-like, with four ambiguous points in between. As for the exposure items, we calculated the total ?dt duration (closure + VOT) for each item: For the Male voice, the selected ambiguous range was 20–116 ms (average: 65.2 ms); for the Female voice, the range was 34–140 ms (average: 96 ms) ($t(1.6) = 1.38$, $p = .19$). The spectral ranges of the two voices were also similar to one another, with no significant difference in the means across the test items (for frequency: Male average = 4564.1, Female average = 4907; $t(1,6) = .5$, $p = .62$; for RMS amplitude: Male = .09, Female = .1, $t(1,6) = .7$, $p = .47$).

Items were presented in a random order, and blocked by voice. The order of voice was counterbalanced. Again, this enables us to look at the question of whether participants characterize items on the continua based on the most recently heard pronunciation.

Procedure

Participants were randomly assigned to one of the lexical decision conditions. Up to three participants were tested simultaneously in a soundproof booth. In the lexical decision task, participants were instructed to respond 'Word' or 'Non-word' to each item by pressing the corresponding button on a response panel. For the experimental groups, items were blocked by voice (Male or Female), and the order of presentation was counterbalanced. Participants were not told that some of the items might have ambiguous sounds.

After the lexical decision phase, all participants categorized sounds on /d/-/t/ continua presented in both the Male and the Female voices; for control participants, this included the voice they had been exposed to as well as a previously unheard voice. The continua were

[2] We also analyzed each measure (closure, and post-burst aspiration) independently and also found no difference between the Male and Female voices: For closure, $t(1,38) = .7$, $p = .48$ (23.3 vs. 21 ms for Male and Female items, respectively); for aspiration, $t(1,38) = 1.61$, $p = .12$ (43 vs. 48.5 ms).

blocked by voice. The order of presentation voice was counterbalanced. Ten randomizations of each continuum were presented.

Results and discussion

Lexical decision

Any participant whose accuracy on the lexical decision task was below 70% was replaced. Nine of the 48 experimental participants were replaced for this reason, and 11 of the 64 control participants were replaced. Table 1 provides the accuracy and reaction time (RT) data for each type of critical item (ambiguous ?D or ?T vs. natural /d/ or /t/) for participants in the experimental and the control conditions.

Experimental and control participants performed similarly on the lexical decision task. Both groups had mean accuracy of over 90% (for the experimental participants, mean accuracy was 93.7%; for the control participants, it was 92.4%). In both groups, accuracy was significantly higher for the natural versions of the critical items than for the ambiguous versions. The experimental groups had a mean accuracy of 97.9% for the natural versions, and 89% for the ambiguous versions ($F1(1,95) = 37.645$, $p < .001$; $F2(1,9) = 17.684$, $p = 0.002$, $minF\ (1,19) = 12.03$, $p < .01$). The control groups had a mean accuracy of 97.2% for natural versions, compared to 88.1% for ambiguous versions ($F1(1,63) = 23.962$, $p < .001$; $F2(1,9) = 17.361$, $p = .002$, $minF\ (1, 25) = 10.06$, $p < .01$).

In addition, participants in both groups correctly labeled ambiguous items as words more quickly than they labeled natural items, suggesting that the lowered accuracy for such items might be the result of a speed-accuracy tradeoff. The difference in response times to ambiguous vs. unambiguous items was significant for the experimental groups (851 vs. 895 ms, respectively; $F1(1,95) = 8.094$, $p = .005$; $F2(1,9) = 5.165$, $p = .049$; $minF\ (1,25) = 3.32$, $p = .08$). For the control conditions, the difference was not significant (850 vs. 880 ms; $F1(1,63) = 2.561$, $p = .12$; $F2(1,9) = 2.607$, $p = 0.14$; $minF \cup (1,32) = 1.21$, $p = .26$). These results are quite consistent with previous perceptual learning results (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2006).

Overall, these data ensure that any perceptual learning differences that we might see between the two groups are not due to their performance (or perception) in the lexical decision phase. Instead, any subsequent differences on the category identification task would have to be attributed to the fact that the experimental group heard a second (and opposite) ambiguous pronunciation during Phase I, while the control group did not.

Category identification

For each participant, we calculated the average percentage of test syllables identified as /d/. For the control groups, there was a clear effect of lexical decision exposure condition on phonemic categorization performance. Listeners who were exposed to the ambiguous ?dt phone in words that normally have a /d/ categorized more items on our continua as /d/ (39.4%) than those who heard ?dt in words which normally have a /t/ (35.0%), $F(1,120) = 5.33$, $p = .02$. Further, this training effect did not interact with voice at test ($F(1,120) = 0.046$, $p = .83$), demonstrating that the size of the effect was just as large across voices (4.8%) as it was within-voice (4.0%).

The significant training effect for the control conditions confirms that people do use lexical knowledge to adjust their perceptual representations, even for categorically perceived stop consonants. Further, it establishes that hearing only 10 critical (ambiguous) items is sufficient to induce perceptual learning. The results here closely replicate the finding in Kraljic and Samuel (2006) that perceptual learning generalizes to new voices for these stop-based stimuli. In that study, the across-voice shift (3.9%) and the within-voice shift (3.5%) were based on twice as many critical items, but the effects are clearly very similar to those generated with only ten. For

Table 1
Experiment 1, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?dt/	
	/t/	/d/	?T	?D
Experimental				
% Correct	97.3%	98.5%	97.9%	80%
RT (in ms)	894	896	874	828
Control				
% Correct	96.6%	97.8%	96.9%	78.5%
RT (in ms)	849	912	873	827

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words, experimental and control conditions.

comparison purposes, Fig. 1 shows the labeling functions for the effect we obtained in the present experiment, using only 10 critical items (left panel), as well as the labeling functions for the effect we obtained (and reported in Kraljic & Samuel, 2006) using 20 critical items (right panel).

Given the reliable effects for the controls, we can now ask whether these effects are preserved when a listener has heard different voice-specific "interpretations" of the ambiguous ?dt. If perceptual learning is speaker-specific, then conflicting information in a very different voice must be irrelevant. Note that the experimental analyses *must* be restricted to the within-voice cases to be meaningful, because participants had initially been trained in *both* voices, but with opposite pronunciations for each voice. Therefore, we must rely on the within-voice test as a measure of whether there was perceptual learning that was consistent with training. The results are clear: Despite the reliable shifts found for the controls, no perceptual learning was observed for listeners who had heard conflicting stimuli in different voices. Participants who had been exposed to the ambiguous sound in /d/ contexts (in either voice) did not categorize more items on the voice-consistent continuum as /d/ (40.3%) than those who learned it in /t/ contexts (38.6%), $F(1,92) = 0.562$, $p = .46$ (see Fig. 2).

When we separate the voices, there is a perceptual learning trend in the male voice (of about 4.0%), but not in the female voice (where the shift is 0.7% in the wrong direction). However, as expected, neither shift was significant for the male data: $F(1,46) = 2.518$, $p = .119$; for the female data $F(1,46) = 0.034$, $p = .855$.

The data show, then, that hearing the same sound as a /d/ in one voice and then as a /t/ in a different voice (or vice versa) results in a net lack of a perceptual learning effect (experimental conditions); hearing only one of these pronunciations results in perceptual learning (control conditions).

Why is there no perceptual learning after hearing a second pronunciation? It might be the case that hearing the first pronunciation shifts perception in one direction, and the effect of hearing the second pronunciation is to shift perception back to 'baseline', so that testing reveals no shift. Another (and theoretically quite different) possibility is that hearing the first pronunciation shifts perception in one direction, and the effect of hearing a new voice at training causes perception to start back at baseline, so that the next pronunciation can be adjusted to in the appropriate way. To separate these two possibilities, we analyzed separately cases in which testing occurred *immediately* after exposure to a particular pronunciation, versus those in which testing occurred later, after *intervening* exposure to the other pronunciation (recall that we used a blocked design at exposure specifically for this purpose). If hearing opposite pronunciations simply serves to cancel one another out, we should not see perceptual learning for either the immediate or the intervening cases. If, however, hearing a new voice during exposure is accompanied by a return to baseline, we should see perceptual learning in the immediate case, where such learning is consistent with the most recent pronunciation heard.

When we include recency of training as a factor, we find a significant interaction between training (?D or ?T) and recency (immediate vs. intervening); $F(1,88) = 5.536$, $p = .02$. As Table 2 shows, when participants are tested immediately after training on a particular sound, they show perceptual learning shifts of about 6.7% in the direction that is consistent with the training they have just had. But, when participants are tested after an intervening training block (with the opposite pronunciation), they show *negative* perceptual learn-

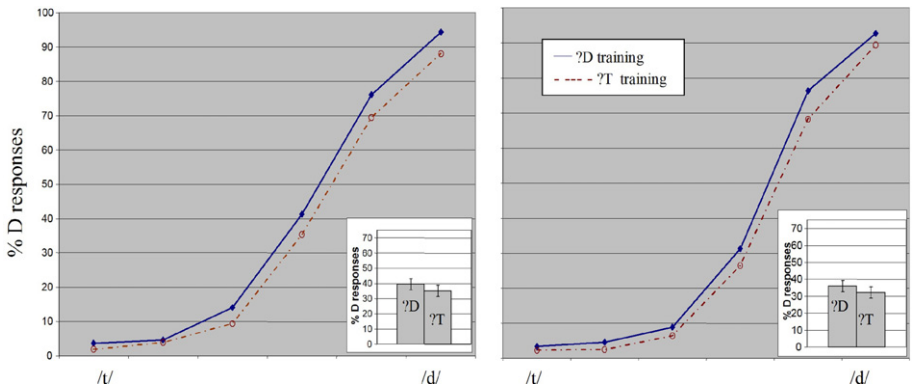

Fig. 1. Experiment 1, The percentage of "D" responses to each item on the test continua for Experiment 1's control group (left panel) and for the participants in Kraljic and Samuel, 2006 (right panel). The perceptual learning effect for Experiment 1's control group, who were exposed to only 10 critical items was remarkably comparable to effects we have found in previous experiments, using 20 critical items. Inset panel shows the same data with 95% confidence intervals for each mean.

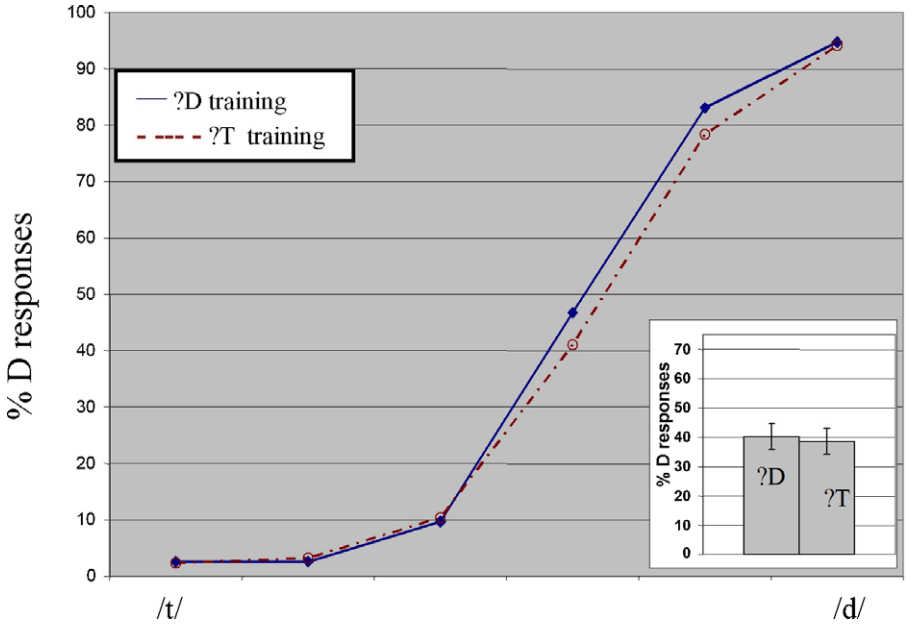

Fig. 2. Experiment 1, The percentage of "D" responses to each item on the test continua. Data points include each participant's responses to test the test continuum that was in the same voice as the relevant training stimuli. These participants demonstrated no perceptual learning after being exposed to two voices, each with a different pronunciation of the critical sound, suggesting that for stop consonants, the perceptual system is unable to maintain multiple, speaker-specific phonemic representations. Inset panel shows the same data with 95% confidence intervals for each mean.

Table 2
Experiment 1, categorization task performance

Training order	Test voice immediate condition (%)	Test voice intervening condition (%)
Male ?D + Female ?T or Male ?T + Female ?D	Female 4.40	Male −1.08
Female ?D + Male ?T or Female ?T + Male ?D	Male 9.00	Female −5.75
Average	6.70	−3.42

Size and direction of the perceptual learning shift, in the appropriate voice, as a function of recency of training. A positive shift indicates perceptual learning consistent with that voice's training; a negative shift indicates perceptual learning that is inconsistent with that voice's training.

ing shifts of about 3.4%: That is, the shifts are 3.4% in the direction that is consistent with the intervening training sound and voice, not the original sound and voice they are being tested on. The perceptual learning effect-by-recency interaction is significant for the male test voice (where the shift is approximately 9.0% for immediate testing, versus −1.1% for intervening; $F(1,44) = 4.671$, $p = .036$), and not quite significant for the female voice (immediate shifts of 4.4% versus intervening shifts of −5.8%; $F(1,44) = 1.976$, $p = .16$). The Male immediate shift of 9.0% is significant $F(1,22) = 7.093$, $p = .014$, but the female immediate shift of 4.4% is not ($F1,22) = 0.588$, $p = .451$. Note that the number of participants in each of these cells is quite small, however; when we collapse over voices, the immediate shift of 6.7% is significant ($F1,44) = 4.04$, $p = .05$.

Overall, these data suggest that, at least for the timing-contrast in our stop consonants, the most recently encountered pronunciation overrides previous exposure. This leads to several implications about the way the perceptual system adjusts to multiple speakers. It is clear that when the critical dimension to be learned (in the present study, the temporal structure of a stop consonant) does not provide local acoustic information about the speaker, the critical phoneme is not adjusted to in a speaker-specific way; instead, it seems that a single representation is shifted each time a new pronunciation is encountered. The effect of hearing a new speaker might

be to 'reset' the representation to baseline, so that it can be appropriately adjusted for the next pronunciation. In Experiment 2, we examine whether the system handles spectral variation in fricatives in the same way. Such variation provides more information about speaker identity, and may therefore produce a more speaker-specific pattern of perceptual learning.

Experiment 2

Experiment 2 was identical to Experiment 1, except that the critical phonemes were fricatives (/s/ and /ʃ/) rather than stop consonants. Given that perceptual learning appears to be largely speaker-specific for fricatives (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), we expect that using this spectral contrast in the current design could result in quite different adjustments than those in Experiment 1. Specifically, if listeners are able to maintain speaker-specific representations, then what has been learned for one speaker should not be affected by subsequent pronunciations from a different speaker.

Method

Participants

Forty-eight undergraduate psychology students from the State University of New York at Stony Brook participated for research credit or for payment. All participants were 18 years of age or older, and all were native English speakers with normal hearing. None had participated in any of the previous experiments.

Design

The design was identical to Experiment 1's experimental conditions: participants were assigned to one of two between-subject Exposure conditions (Male ?S + Female ?SH or Male ?SH + Female ?S) in which they performed an auditory lexical decision task. Exposure was blocked by voice, and order of the voices was counterbalanced. Based on the very large effects that we have found with these fricative stimuli (e.g., Kraljic & Samuel, 2005), we expected to find significant perceptual learning effects in this experiment. Given this, and the successful results of the control condition of Experiment 1, we did not include the control conditions here.

The Exposure phase was followed by a within-subject Categorization phase. All participants categorized items on /asi/-/aʃi/ continua in both the male and the female voices.

Materials and procedure

The procedure was identical to Experiment 1's experimental conditions. The only difference was that the critical phonemes in the present experiment are /s/ and /ʃ/, not /d/ and /t/.

Phase 1: Exposure (lexical decision). Two experimental lists were created for an auditory lexical decision task, each with 100 words and 100 nonwords (a complete list of the words used can be found in Kraljic & Samuel, 2005). The lists were identical except for the 40 critical words, 20 of which contained the critical /s/ phoneme (but no /ʃ/), and 20 of which contained /ʃ/ (but no /s/). As in Experiment 1's experimental conditions, each subject heard all 40 critical items, but half (10 /s/ and 10 /ʃ/) were presented in a Male voice, and the other half were presented in a Female voice; items were counterbalanced across subjects.

The same criteria were used to select the critical and filler words as in Experiment 1: words ranged in length from two (*brochure*, *obscene*) to four (*negotiate*, *hallucinate*) syllables, the two sets of critical words were matched to each other and to the filler words in terms of mean number of syllables and mean frequency, and the filler words could not have any instance of the critical phonemes /s/ or /ʃ/. The procedure for creating the nonwords was identical to that in Experiment 1.

Details about the recording procedure and stimulus construction can be found in Kraljic and Samuel (2005). The acoustic properties of /s/ and /ʃ/ enable a fairly straightforward mixing of the two sounds to create an ambiguous ?sʃ mixture for each critical item. /s/ and /ʃ/ are very similar in both duration and amplitude; the main difference between the two is in frequency - /s/ is higher frequency than /ʃ/. Thus, the /s/ and /ʃ/ for each item were simply mixed together in various proportions, until a mixture was created that both authors identified as ambiguous. In this way a single ambiguous mixture for each critical item was selected for use in the experiment.

We calculated the spectral means for the fricative 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 (we used Praat) to obtain a single number for each item. This number represents the mean frequency of the excised portion's spectrum. As expected (and in contrast to Experiment 1's temporal contrast), the spectral means varied systematically across the male and female voices (4935 vs. 5450 Hz, respectively, $t(1,38) = 2.87$, $p = .004$).

We also calculated the duration for each critical fricative. Our measurements confirmed that the duration of the fricatives did not vary systematically across the male

and female voices (154.5 vs. 163.8 ms, respectively, $t(1,38) = 1.64$, $p = .10$).

Participants in the lexical decision task were randomly assigned to one of two groups, Female?S + Male?SH or Female ?SH + Male?S. The order of voices was included as a factor for counterbalancing. The instructions, procedures, and response recordings were identical to Experiment 1.

Phase II: Test (category identification). In the test phase of the experiment, all participants categorized six sounds on two separate /asi/-/aʃi/ continua, one in each voice they had heard during exposure Spectral means for the Male voice ranged from 3963 to 5968 Hz (average: 4943 Hz); those for the Female voice ranged from 5275–6784 Hz (average: 6099 Hz). The procedure for creating the continua was the same as that for creating the ambiguous critical items used in the lexical decision task. The continua were blocked by voice (Male and Female), and the order of presentation voice was counterbalanced across participants. Participants responded by pressing a button labeled "S" if the sound they heard sounded like an /s/, or pressing a button labeled "SH" if it sounded like an /ʃ/. Ten randomizations of the six sounds on the /asi/-/aʃi/ continuum were presented in one voice; then participants performed the same categorization task on the /asi/-/aʃi/ continuum in the other voice. Responses and response times were recorded.

Results and discussion

Lexical decision

Any participant whose accuracy on the lexical decision task was below 70% was replaced. One of the 48 participants had to be replaced for this reason.

Table 3 provides the accuracy and RT data for each type of critical item (ambiguous ?S or ?SH versus natural /s/ or /ʃ/). Overall, participants performed very well on the lexical decision task; mean accuracy was 96.9%. Accuracy for unambiguous items (99.2%) was somewhat higher than accuracy for ambiguous items (94.8%), $F1(1, 95) = 25.07$, $p < .001$; $F2(1,9) = 19.456$, $p = .002$; min $F'(1,27) = 10.95$, $p < .01$. Participants were also slightly faster to correctly label unambiguous items as words (964 ms) than to correctly label ambiguous items (995 ms). This difference was marginally significant both by-subjects ($F1(1,95) = 3.349$, $p = .07$) and by-items ($F2(1,9) = 3.527$, $p = .093$); *min* $F'(1,34) = 1.72$, $p = .2$. These data indicate that our ambiguous mixtures were relatively natural sounding to our participants.

Category identification

For each participant, we calculated the average percentage of test syllables identified as /ʃ/. Unlike the experimental groups in Experiment 1, participants in the present experiment demonstrated very strong voice-consistent perceptual learning: Participants who were exposed to ambiguous /ʃ/ in a particular voice subsequently categorized significantly more items in that voice as /ʃ/ (67.4%) than did participants who had been exposed to ambiguous /s/ in the same voice (49.1%) ($F(1,92) = 50.63$, $p < .001$; see Fig. 3).

This perceptual learning effect did not interact with training voice ($F(1,92) = .409$, $p = .524$), indicating that the effect was just as strong for the Male voice (which produced a shift of 16.7%) as it was for the female voice (where the shift was 19.9%). Both the male effect and the female effect were highly significant ($F(1,46) = 17.032$, $p < .001$ and $F(1,46) = 39.109$, $p < .001$, respectively).

In contrast to the results for Experiment 1's stop consonants, there was no interaction between the perceptual learning effect and recency of exposure (i.e., testing Immediately versus testing after an Intervening, and opposite, exposure), $F(1,88) = 0.193$, $p = .662$. Table 4 presents this breakdown of the data. The effect for an immediate test was 17.6% ($F(1,46) = 22.17$, $p < .001$), while after intervening training, the effect was 19.4% ($F(1,46) = 28.14$, $p < .001$). Clearly, then, being exposed to a different speaker with a different pronunciation in between the original exposure and the test did not attenuate the effect. This finding reinforces the idea that perceptual experience leads to very different learning for the spectral variation in fricatives than it does for temporal variation in stop consonants.

Table 3
Experiment 2, lexical decision task performance

	Critical words			
	Natural		Ambiguous /?sʃ/	
	/s/	/ʃ/	?S	?SH
% Correct	98.9%	99.4%	93.1%	96.5%
RT (in ms)	964	964	1037	953

Mean accuracy and reaction times (for correct items) for natural and ambiguous critical words, both Experimental groups.

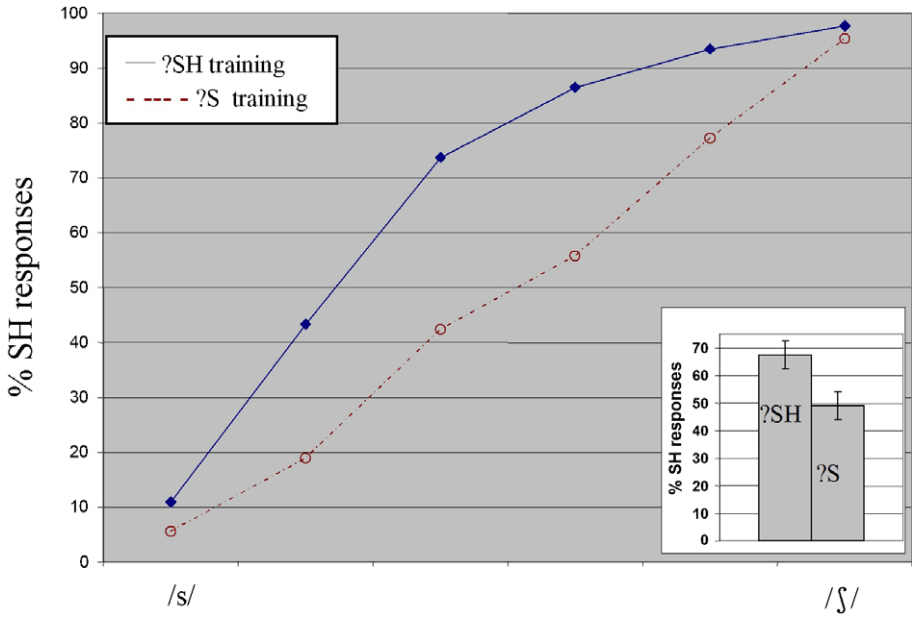

Fig. 3. Experiment 2, The percentage of "SH" responses to each item on the test continua for all participants in the experimental conditions. Data points include each participant's responses to test the test continuum that was in the same voice as the relevant training stimuli. In contrast to what we found with stop consonants, listeners *do* appear to be able to maintain multiple representations for fricatives, as evidenced by this large, speaker-specific perceptual learning effect. Inset panel shows the same data with 95% confidence intervals for each mean.

Table 4
Experiment 2, categorization task performance

Training order	Test voice immediate condition (%)	Test voice intervening condition (%)
Male ?S + Female ?SH or Male ?SH + Female ?S	Female 20.6	Male 19.7
Female ?S + Male ?SH or Female ?SH + Male ?S	Male 13.7	Female 19.2
Average	17.6%	19.4%

Size and direction of the perceptual learning shift, in the appropriate voice, as a function of recency of training. A positive shift indicates perceptual learning consistent with that voice's training; a negative shift indicates perceptual learning that is inconsistent with that voice's training.

General discussion

The data from Experiments 1 and 2 show an intriguing pattern, and demonstrate that the nature of the phonemic representations themselves will determine the nature of learning that the representations can support. Previous studies have shown that adjustments to 'odd' pronunciations in stop consonants generalize to new speakers, while adjustments to fricatives are often speaker-specific (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006). The present experiment confirms those findings and extends them to situations in which the system has to adapt to different speakers who produce different pronunciations of the same sound. When the critical sound varies along a temporal-voicing dimension (Experiment 1's stop consonants), hearing a new voice appears to serve as a 'resetting' cue, to return to the standard representation in preparation for adapting to the current speaker. The result is that categorization of the ambiguous sound will be consistent with the most recent speaker heard, as a priming account predicts (Dell & Brown, 1991; Pickering & Garrod, 2004). But when the critical sound varies along a spectral-place dimension (Experiment 2's fricatives), the system appears able to maintain multiple representations simultaneously, each for the appropriate speaker.

In the introduction, we speculated that these differences might be explained by supposing that the perceptual system adjusts to different pronunciations

and different speakers by using purely acoustic cues. When those cues simultaneously provide information about the speaker's identity (as they do with the critical fricatives in Experiment 2), learning reflects speaker-based adjustments. But when the acoustic cues are uninformative with respect to speaker (as they are for the stop consonants in Experiment 1), learning is independent of speaker. Acoustic analyses of our stimuli confirm that, as we have argued, our fricatives provided speaker specific information while our stop consonants did not: The spectral means for Experiment 2's critical (ambiguous) fricatives varied systematically across the male and female voices, while the temporal cues for Experiment 1's critical (ambiguous) stops showed no such difference across the male and female voices.

The presence of speaker-specific cues in our fricatives, and the lack of such cues in our stops, is in fact entirely consistent with how these cues naturally pattern. In English, the fricatives /s/ and /ʃ/ are always distinguished from one another within-speaker by the spectral mean: /s/ is higher in frequency than /ʃ/ (see, e.g., Jongman et al., 2000; McFarland, Baum, & Chabot, 1996; Nittrouer, Studdert-Kennedy, & McGowan, 1989). Critically, this variation in spectral cue to phoneme is correlated with (at least) the gender of the speaker: Females' productions of both /s/ and /ʃ/ are higher in frequency than males' productions (Jongman et al., 2000). Further, although raw productions of /s/ and /ʃ/ may vary quite a bit across talkers, it seems that listeners are sensitive to the consistency with which any one talker produces those sounds (e.g., Newman, Clouse, & Burnham, 2001).

While listeners may also be sensitive to talker differences in the voice onset time (VOT) of stop consonants (e.g., Allen & Miller, 2004), it is far less clear that VOT varies in a way that can be predicted by speaker (or speaker gender); in fact, a recent study (Allen, Miller, & DeSteno, 2003) finds that the biggest predictor of VOT is speaking rate (both within and across speakers), and not speaker. Given that variation in VOT is far less systematically predicted by speakers than variation in spectral mean is, it makes sense that listeners would make adjustments that encode speaker-specific information for fricatives, but not for stops.

This raises another interesting question, which we hope to resolve in future research: Is the nature of perceptual learning (speaker-specific versus more general) determined on a case-by-case basis, depending on what the current stimuli afford (as we have in fact been assuming), or does learning reflect more general tendencies in the language? That is, if listeners were exposed to stop consonants that varied more systematically with respect to speaker, could the perceptual system take advantage of that systematicity and learn in a speaker-specific way?

The present stimuli were not designed to address this question, because they reflect the real-world tendency for stops and fricatives to pattern differently with respect to speaker-specificity. It is certainly the case that different classes of sounds are in general handled differently by the perceptual system; some studies have found very fine-grained distinctions in the processing of different sounds, and have suggested that the cues in the acoustic signal can actually determine the mechanism by which the sound is processed (e.g., whether the sound is processed by a central mechanism that operates at a cognitive level, or by a peripheral mechanism that operates on the acoustic signal; for a discussion see Jamieson & Cheesman, 1986). Further, fricative categories appear to be more malleable both productively and perceptually than stop consonants are, as well as more difficult for children to establish (e.g., McFarland et al., 1996; Nittrouer, 1995; Nittrouer et al., 1989). The literature thus suggests that the perceptual learning differences we have found are the result of general tendencies in the processing of stops versus fricatives. However, the fact that listeners do seem to be sensitive, at least under certain conditions, to individual talker differences in VOT (Allen & Miller, 2004), as well as our own findings that perceptual learning of fricatives will generalize to a new speaker if the spectral means are ambiguous (Kraljic & Samuel, 2005), argue for a flexible perceptual system, one that adjusts on the basis of current affordances, and not only general tendencies.

Finally, we should point out that we cannot yet specify which of the differences in our stimuli are the primary cause of the differences we find in processing. We have discussed two distinctions between our stop and fricative contrasts: (1) the presence (fricatives) or absence (stops) of local speaker-specific cues, and (2) the primarily temporal variation in stops versus the primarily spectral variation in fricatives. In our experiments, these cues are correlated, as they are in spontaneous speech. As we have noted, either of these may independently be causing listeners to adapt differently in the two experiments. There is at least one other distinction between the stops and fricatives: Our stop contrast differed from one another in voicing, whereas the fricatives differed in place of articulation. If further research ultimately demonstrates that stops and fricatives are processed differently from one another regardless of the presence of speaker-specific information, one or both of the other two distinctions may prove critical for explaining the differences.

The present data and the previous research we have described suggest that perceptual adjustments are the result of quite local, mechanistic processes. But we know that social factors can also be important in shaping the behavior of speakers and listeners, even at the phonemic level. At the simplest level, there is a well-established tendency for people in a conversation to become more alike in terms of linguistic (and non-linguistic) features; such convergence extends

to the pronunciation of consonants and to conversationalists' vowel space (see Coupland, 1984 for a discussion). This tendency is probably not the result of a conscious effort to change one's speech: Goldinger (1998) found that speakers who are shadowing (i.e., repeating) speech presented over headphones tend to spontaneously imitate various acoustic parameters of the voices they are listening to, including the fundamental frequency and word durations. Ongoing work in our own lab suggests that such imitation is present for parameters that participants do not report being aware of, and even have difficulty explicitly reporting. Sancier and Fowler (1997) report a case study of a native speaker of Brazilian Portuguese who traveled between Brazil and the US. The voice onset times (VOT) for this speaker's stop consonants (when she spoke in Portuguese) were shifted following time spent in the US, bringing them closer to the VOT's used in American English.

Thus, while the present study demonstrates that acoustic experience plays a large and probably very automatic role in shaping subsequent perception, we are not suggesting that it tells the whole story. Many other studies suggest that any complete model of speech perception will have to include a role for less automatic and less purely acoustic (though not completely conscious and deliberate) processes, such as changes in pronunciation that might reflect the language background, social status, or linguistic environment of a particular interlocutor. Ongoing work in our lab is thus beginning to explore the question of perceptual experience from a more social perspective. The goal of this work is to clarify the conditions under which listeners' goals, attributions, and higher-level knowledge might shape the perceptual adjustments that the current study has demonstrated.

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 Susan Brennan and to Donna Kat for their support and feedback, and to Dennis Norris, Jim Sawusch, and one anonymous reviewer for their suggestions.

References

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115*, 3171–3183.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113*, 544–552.

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

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

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

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener". In D. J. Napoli & J. A. Kegl (Eds.), *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman* (pp. 105–129). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

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

Jamieson, D. G., & Cheesman, M. F. (1986). Locus of selective adaptation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 286–294.

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

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

McFarland, D. H., Baum, S. R., & Chabot, C. (1996). Speech compensation to structural modifications of the oral cavity. *Journal of the Acoustical Society of America, 100*, 1093–1104.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects in the comprehension of referring expressions. *Journal of Memory and Language, 49*, 201–213.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329–336.

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

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *Journal of Phonetics, 97*, 520–530.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech, Language and Hearing Research, 32*, 120–132.

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

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Brain and Behavior Sciences, 27*, 169–190.

Sancier, M., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics, 25*, 421–436.

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