Attention, Perception, & Psychophysics
2009, 71 (6), 1207-1218
doi:10.3758/APP.71.6.1207

TUTORIAL REVIEW

Perceptual learning for speech

ARTHUR G. SAMUEL
State University of New York at Stony Brook, Stony Brook, New York

AND

TANYA KRALJIC
University of Pennsylvania, Philadelphia, Pennsylvania

Adult language users have an enormous amount of experience with speech in their native language. As a result, they have very well-developed processes for categorizing the sounds of speech that they hear. Despite this very high level of experience, recent research has shown that listeners are capable of redeveloping their speech categorization to bring it into alignment with new variation in their speech input. This reorganization of phonetic space is a type of perceptual learning, or recalibration, of speech processes. In this article, we review several recent lines of research on perceptual learning for speech.

Perception has long intrigued those among us who seek to understand how our minds and brains make sense of the world. Perception, far from being an objective translation of reality, is shaped both by a perceiver's knowledge and by his or her past experience with particular stimuli. A century and a half ago, Volkmann (1858) showed that even the basic perception of touch (whether one feels a single touch or two) is subject to experience: With practice, subjects resolved smaller and smaller physical separations as two separate touches. A century later, J. J. and E. J. Gibson published a seminal article on such "perceptual learning" (Gibson & Gibson, 1955). They defined perceptual learning as making the perceiver "more sensitive to the variables of the stimulus array" (p. 40), and their work provided a framework for a (quite varied) subfield of experimental psychology.

Many years and many experiments later, Goldstone (1998) published a review of perceptual learning in which he developed a somewhat more nuanced definition. He defined perceptual learning as "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment" (p. 586). This definition includes a number of key points (e.g., that the changes are long lasting, and that they allow the organism to adapt to the prevailing environment) that will be useful when we consider recent work on perceptual learning for speech, the focus of our review.

Goldstone (1998) suggested that perceptual learning could be produced by four different mechanisms. Conceptually, perhaps the simplest is "stimulus imprinting"—the notion that the perceptual system essentially grows a detector of some sort that is specifically tuned to the environmentally relevant stimulus. This detector could theoretically represent any level of abstraction, including basic features (e.g., a line at a particular orientation) as well as more complex patterns or even an entire previously experienced stimulus (e.g., a particular chair). Two other mechanisms are complementary: "differentiation" and "unitization." In the former, the perceptual system develops the ability to separate critical parts from the whole (e.g., being able to detect the small stroke in "Q" that differentiates it from "O"). In the latter, the system learns to collect parts together into a more functional whole (e.g., a skilled driver can see the flow of activity of cars, bicycles, and pedestrians, rather than merely registering the individual motion of each object). The final mechanism, "attentional weighting," is based on the notion that with experience, more weight can be given to relevant features of a stimulus. For example, a native speaker of English will pay little attention to a vowel's length, since length is an irrelevant feature in vowel identification in English (Strange, Jenkins, & Johnson, 1983).

At the time of Goldstone's (1998) review, most perceptual learning research had been focused on visual processing. But in the last decade, there has been an explosion of research that examines perceptual learning for speech, offering a complementary perspective and, ultimately, a more elaborated understanding of how perceptual learning might work. Work on perceptual learning for speech has developed into several different literatures that focus on different aspects of this topic. Across these literatures,

A. G. Samuel, asamuel@ms.cc.sunysb.edu

the commonality is that listeners are exposed to speech that is in some way noncanonical, or different than the speech usually experienced, and this exposure produces a change in subsequent spoken language processing.

In the following review, we sort the literature into two themes. In Theme I, the procedures and results clearly fit both Goldstone's (1998) and Gibson and Gibson's (1955) definitions of perceptual learning: Listeners are given experience with some kind of unfamiliar speech stimuli, and the exposure leads to improvement in their ability to identify or discriminate speech stimuli of that type. Three types of unfamiliar speech are considered: phonetic contrasts in a nonnative language; accented or dialectal speech; and degraded speech (e.g., through compression or noise). In all of these studies, perceptual learning is inferred when exposure to the challenging speech leads to a better ability to understand what is being said.

The research in Theme II has appeared quite recently. In these studies, the perceptual learning effect is consistent with Goldstone's (1998) definition, but may not meet Gibson and Gibson's (1955) standard, because the learning allows the listener to be more attuned to the prevailing environment, but not necessarily to make finer discriminations than before. The basic procedure is to present listeners with phonetically ambiguous stimuli, with some source of contextual information to disambiguate the stimulus. Perceptual learning is indexed by a shift in phonetic categorization. Presumably retuning should help the listener categorize speech better in the prevailing input environment, but the measurement is most often of category boundary location rather than of comprehension. A great virtue of this work is that the observed category boundary shifts provide a clear indication of exactly what is changing in perceptual processing as a function of experience, something that is generally not available for the Theme I studies.

THEME I

Perceptual Learning That Improves Perception of Difficult Speech

I.1. Perceptual Learning of Nonnative Phonetic Contrasts

The most thoroughly studied case of perceptual learning of a nonnative phonetic contrast involves the discrimination of English /r/ and /l/ by native speakers of Japanese. This contrast does not exist in Japanese—there is a single flap that is not a very good match to either /r/ or /l/, and native Japanese speakers are known to have difficulty with this distinction. According to Aoyama, Flege, Guion, Akahane-Yamada, and Yamada (2004; see also Cutler, Weber, & Otake, 2006), the Japanese segment is phonetically somewhat closer to English /l/ than to English /r/, but again, neither English sound is a good match. Logan, Lively, and Pisoni (1991; Lively, Logan, & Pisoni, 1993) conducted a set of training experiments in which they gave native Japanese speakers (who were living in the United States) three weeks of training with the /r/–/l/ contrast. In most of the conditions, the listeners heard English words with /r/ or /l/ produced by five different talkers, in a variety of different phonetic contexts (e.g., word-initial; in a cluster such as /pl/; word-final). The listeners were asked to decide whether a given stimulus included /r/ or /l/, and they received feedback on each trial. The authors found that these high-variability learning conditions did indeed produce significant, if somewhat modest, improvements in how well the listeners could identify /r/ versus /l/. Critically, the learning generalized to new tokens, from different talkers. In contrast, if the training was with a single talker, the learning only generalized to new tokens, not to other talkers.

Lively, Pisoni, Yamada, Tohkura, and Yamada (1994) used a similar procedure, but trained monolingual Japanese listeners living in Japan. Presumably because these subjects had less prior exposure to the /r/–/l/ contrast than did the subjects in the prior studies (who were living in the United States), the improvement through training was larger. An important addition in this study was the inclusion of three-month and six-month posttraining retests. Despite the absence of any intervening training, the improvements were still largely intact after three months; they were reduced but still present after six months, consistent with Goldstone's (1998) criterion that perceptual learning is "relatively long-lasting." An interesting feature of these studies is that even when there is generalization to new talkers, there are usually bigger effects for test words produced by a voice used in training, even months later. This difference reflects an inherent tension that must be resolved in a learning system: Specifically tuning the representations/processes for the training input may optimize performance on the stimuli prevailing in the environment, but there are also advantages to being able to generalize the learning to new sources.

Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999) tested whether the gains made with perceptual training also enhance the ability of Japanese speakers to *produce* the English /r/–/l/ distinction. Bradlow et al. (1997; Bradlow et al., 1999) used training procedures similar to those of Lively et al. (1994), and measured both perceptual and production accuracy before and after the training regime. Production accuracy was judged by a group of naive American listeners. In addition to perceptual improvements, Bradlow et al. (1997; Bradlow et al., 1999) found moderate but significant improvements in production. Like the perceptual gains, those in production were intact after a three-month period with no additional training. The authors interpreted the transfer from perception to production as evidence for shared representations across the two systems. Whereas the transfer found by Bradlow and colleagues might reflect such a tight perception–production link, it could alternatively simply reflect the acoustic availability of articulatory targets (i.e., after perceptual training, the subjects are able to hear what they need to try to produce). This is indeed a link between perception and production, but one that is weaker and does not imply shared representations.

The perceptual learning of foreign language sounds has been investigated in other languages and with other

types of phonetic distinctions as well. Wang, Spence, Jongman, and Sereno (1999) applied Logan et al.'s (1991) high-variance training approach to native English speakers trying to learn Mandarin tones. In Mandarin, there are four different prosodic patterns that can be imposed on a syllable, and these four tones combine with the phonetic information to determine the lexical identity of the word. A given syllable has one meaning with one tone, and a quite different meaning with a different tone. Native English speakers have great trouble with these tonal distinctions, just as native Japanese speakers have trouble with /r/ and /l/. Wang et al. gave students in the United States (who were taking college Mandarin courses) two weeks of training on the tone distinctions, and found substantial improvements. In two-alternative forced choice tests, students' accuracy in identifying the tones increased from about 70%, before training, to about 90%, afterward. This improvement was even larger than the improvements found in the /r/–/l/ studies, indicating that perceptual learning of suprasegmental cues can be at least as good as learning of segmental information. Moreover, on a test conducted six months after the training ended, there was very good retention of the learning.

Flege's (1995) study also involved English–Mandarin distinctions, but his focus was on Chinese speakers acquiring an English contrast. Native Mandarin speakers have trouble hearing the difference between final /t/ and final /d/ in English, when these stops are unreleased. A major cue to voicing in final position is vowel duration (Klatt, 1976; Peterson & Lehiste, 1960), which is not the same as in Mandarin. Flege (1995) compared two different training procedures, to determine whether they differed in their ability to produce perceptual learning. Native Mandarin speakers living in the United States received two weeks of training on this contrast. Half of the subjects did an identification task during training—after each item, they reported whether the final stop was /t/ or /d/. For the other half, each trial consisted of two different words, produced by two different talkers, and the two words either both ended in the same consonant (two /t/s, or two /d/s), or they ended in different consonants (one /t/, one /d/). The task was to report "same" when both words ended in the same consonant, or "different" when they did not. Both methods produced good improvement on a posttest, suggesting that both training situations allow the development of perceptual learning.

A final study in this group was done by Kingston (2003). In this case, the subjects were Americans, and the contrasts they were taught were distinctions among German vowels. Kingston provided a very nice discussion of two leading theories of how people learn second-language distinctions—Best's (1994; Best, McRoberts, & Goodell, 2001) perceptual assimilation model, and Flege's (1991) speech learning model. Kingston noted that both of these models predict that how well someone learns to discriminate a nonnative contrast should depend on the extent to which the contrasting sounds map onto different sounds in the learner's native language. These models both correctly predict the difficulty that Japanese speakers have with /r/–/l/, because there are not contrasting sounds in Japanese that /r/ and /l/ can be mapped to. However, Kingston's data pose problems for both models, because the pattern of learning shown by Americans who were trained on certain German vowel contrasts did not match the predictions. In particular, when certain pairs of vowels differed in the same way (e.g., in vowel height), Kingston found cases in which the pairs were not learned equally well. Since the pairs differed in the same way, they should have been mapped onto native vowels (e.g., a pair differing in vowel height) in the same way, leading to equivalent learning, but they were not.

I.2. Perceptual Learning of Accents and Idiolects

We have all had the experience of encountering someone with a strong foreign accent whose speech is initially quite difficult for us to understand. In most cases, after we have listened to such a speaker for a period of time, their speech becomes more intelligible to us. This phenomenon has been taken as a case of perceptual learning in speech: The input has not changed, but over time the listener becomes more skilled in decoding the speech, much as the listeners discussed in the preceding section improved their perception of nonnative contrasts.

Bradlow and Bent (2008) examined how listeners adapt to accented speech, drawing on some of the procedures developed for studying perceptual learning of nonnative contrasts. In particular, they investigated whether—as in learning nonnative contrasts—a high-variability training regime produces better learning. American listeners received two training sessions in which they heard English sentences that had been produced by native Chinese speakers with strong Chinese accents. Half of the listeners were trained with the high-variability method—they heard sentences produced by five different Chinese speakers with accents of varying strength. The other listeners were trained in a low-variability method—all of the sentences they heard came from a single speaker. Their task was simply to transcribe the sentences they were hearing. Following training, they completed a test phase that also involved transcribing sentences. There was a control condition in which listeners heard nonaccented English sentences. This condition provided a baseline measure of the effect of transcription practice.

The pattern of learning was consistent with the results for learning nonnative contrasts. Listeners who had trained with multiple speakers showed about a 10% improvement over the baseline group. This improvement matched the performance of subjects who were trained on a single speaker and tested on that same speaker. In contrast, subjects who trained on a single speaker and were tested with sentences from a different speaker were no better than subjects in the baseline condition. This pattern indicates that experience with a single speaker's accent is sufficient to improve perception of that speaker's speech, but not the speech of others (even others with similar accents). However, if multiple sources are provided, then the learning generalizes to new speakers. This generalization does have its limits: When the listeners were tested on English sentences with a very different accent (Slovakian), there was no significant advantage for any of the training

conditions, including the multiple-speaker case, over the baseline.

Most of the perceptual learning studies we have discussed, both of accented speech and of nonnative contrasts, have provided extensive and explicit training on the unfamiliar speech. Clarke and Garrett (2004) investigated whether much less extensive exposure could suffice to improve the perception of accented speech. They presented listeners with four blocks of sentences, with each block consisting of four Spanish accented sentences (Experiments 1 and 2) or of six Chinese accented sentences (Experiment 3). Each sentence ended with a monosyllabic word that was not well predicted by the preceding sentence context; the final word was followed by a visual target that either matched the identity of the word or did not. Subjects made a simple yes–no matching judgment. All subjects showed steady and substantial improvements in reaction time for this judgment across the first three blocks of trials (12–18 sentences), with these response times approaching individually determined baseline response times for nonaccented stimuli by the fourth block.

Subjects in a control condition responded to nonaccented stimuli for the first three blocks, and received the accented stimuli only in the fourth block. These controls showed a large increase in response times in the fourth block, unlike experimental subjects who had heard 12–18 accented sentences (by the same speaker) before that block. Clarke and Garrett (2004) noted that this training period (the first three blocks) only lasted about 1 min, and in some additional analyses, showed that a significant improvement can be seen after as few as 2–4 accented sentences. The time scale for this improvement is much, much faster than what is seen in most studies of perceptual learning of nonnative contrasts (e.g., /r/–/l/ for Japanese speakers).

Clarke and Garrett (2004) did not include a long-term test of retention, making it difficult to determine whether the improvements would meet Goldstone's (1998) "long-lasting changes" criterion. But McGarr's (1983) study of deaf speech suggests that learning of accented pronunciations may be retained beyond immediate training. Deaf speech (like foreign-accented speech) contains pronunciations that can initially be difficult for inexperienced speakers to understand. McGarr investigated how listeners adjust to such speech, using a more naturalistic methodology than most, and looked at learning over a very long time scale. McGarr compared the speech identification abilities of individuals who had years of experience listening to the speech produced by deaf talkers with the abilities of novice listeners. She found that the experienced listeners consistently outperformed listeners who had not had prior exposure to deaf speech, across a wide range of contextual conditions (e.g., individual words, words in sentences). Critically, the talkers in this study were unfamiliar to both groups of listeners, indicating that the advantage shown by experienced listeners was not tied to individual talkers. Instead, these listeners had developed perceptual processes that allowed them to overcome some of the severe phonetic and prosodic variation present in speech produced by deaf talkers.

In addition to these speaker-general results reported by McGarr (1983), several of the studies that we have discussed here found evidence for speaker-specific learning effects, with improvements primarily or exclusively when test items were spoken by the same talker who had been heard in the training phase. Nygaard, Sommers, and Pisoni (1994; Nygaard & Pisoni, 1998) took the idea of speaker-specific learning to its logical extreme, and examined whether learning to identify individual talkers affected word recognition; they investigated the learning of idiolects. In the training phase in these experiments, listeners heard speech produced by 10 different talkers (5 male, 5 female), and their training task was to identify each item with a particular talker by pushing one of 10 labeled buttons (Mary, Sue, ... , Bobby, John).

After the training on speaker recognition, subjects were given a task in which speech was presented in noise; the task was to report what the speech was. There was an interesting correlation between the talker recognition results and the speech identification results. The subjects were partitioned into those who were good at the talker recognition task (those who achieved a level of talker recognition above a criterion of about 70% correct), and those who were not. For those subjects who had learned to identify the talkers, performance on the speech identification task was better for items that were produced by the familiar talkers than for items produced by new talkers. For subjects with poor talker recognition, there was no difference in speech identification as a function of whether the speech came from a familiar or unfamiliar talker. These results reinforce the view that listeners develop speaker-specific representations (see Allen & Miller, 2004, for evidence that listeners can learn to make talker-specific phonetic categorizations on the basis of voiced onset time). The link between voice information and word recognition supports models of word recognition that include episodic information, at least at some level (Goldinger, 1998, argued for a lexical site; McQueen, Cutler, & Norris, 2006, favored prelexical), rather than exclusively abstract models.

I.3. Perceptual Learning for Degraded Speech Input

We have seen that listeners are able to learn something about accented speech that allows them to improve their recognition of words produced with that accent, even when the talker is unfamiliar. One way to think about accented speech is that it is a form of degraded input: Abnormal properties (carried over from the speaker's native language) deform the phonetic and prosodic patterns that the listener normally would receive. Indeed, a primary purpose of perceptual learning for speech appears to be to allow listeners to understand spoken language that deviates from the norm. For this reason, a number of researchers have examined perceptual learning by intentionally degrading the speech signal.

One method of degrading the signal has been to use speech compression software to create speech that has very abnormal temporal properties. Dupoux and Green (1997) compressed sentences to either 45% or 38% of their normal duration, a rather severe alteration. Sub-

jects transcribed the speech, and their transcriptions were coded for the number of keywords in each sentence that were correctly reported. The more severe compression produced baseline accuracy of only about 30%, and even in the less severe case listeners reached only about 65% accuracy. Dupoux and Green then gave their listeners 15–20 training sentences, and observed improvements of approximately 10%–15% in keyword report. These results are very similar to what Clarke and Garrett (2004) found in their study of perceptual learning of accents: Experience with a small number of sentences, over the course of about 1 min, is sufficient to produce significant improvement. Dupoux and Green also included conditions in which the training set was briefly interrupted by speech from a different talker, or speech at a different rate, and they found only a transitory dip in keyword report. This suggests that the effect does in fact reflect learning, rather than some short-term adaptation.

Pallier, Sebastián-Gallés, Dupoux, Christophe, and Mehler (1998) used similar compression rates (40% of normal) and duration of training (8–10 sentences), but tested both in the same language and in a different language. They did so in order to provide a more detailed assessment of which segmental or suprasegmental patterns are being learned. Listeners were given experience with compressed speech in one language, and then tested with compressed speech in the same language, or in a different language. Pallier et al. compared learning in bilingual listeners who knew both languages with learning in monolingual listeners who knew only the language used in the test sentences; the latter could not understand the compressed training sentences, but had gotten experience listening to compressed speech. Interestingly, knowledge of the training sentences' language did not make much difference in the results. Instead, what seemed to control the amount of transfer was the phonological similarity of the language pair: There was good transfer from Catalan training to Spanish test sentences (two closely related languages), but there was little transfer between English and French (relatively unrelated), whereas Dutch/English (moderately related) produced intermediate results. The authors suggested that when languages share stress patterns and similar syllabic structure and pacing, then what is learned with one language will help with the other; without such shared phonological properties, there is no generalization across the languages. These results provide some of the first evidence about where in the system perceptual learning for speech takes place—at the phonological level. Sebastián-Gallés, Dupoux, Costa, and Mehler (2000) continued the approach of using cross-language training/testing, and reported that the shared phonological properties needed for cross-language efficacy go beyond simple rhythmic structure, probably including the lexical stress pattern and vowel inventories of the languages.

Davis, Johnsrude, Hervais-Adelman, Taylor, and McGettigan (2005) used a different method to degrade the speech input: noise vocoding. In vocoded speech, the speech is divided into a number of frequency ranges, and the time-varying amplitude is extracted for each frequency range. Then, band-pass noise is modulated by the amplitude contour within each frequency range, and the results are recombined. The result is a stimulus that has a gross preservation of the energy distribution of the original speech, but with most of the phonetic details eliminated. If many frequency bands are used, the speech remains intelligible, but with fewer bands it is quite degraded.

Davis et al. (2005) used six frequency bands, a rather severe degradation: Listeners typically can identify fewer than 10% of the words in sentences constructed this way, when they first encounter the vocoded speech. Davis et al. trained listeners with such stimuli, under several different conditions. One manipulation compared learning when listeners heard a clear version of each vocoded sentence in the order vocoded–clear–vocoded, versus vocoded–vocoded–clear; listeners were asked to transcribe as much of each vocoded sentence as they could. In several experiments, they showed that hearing a clear version prior to a vocoded version greatly improved learning (as measured on a transcription posttest without any clear sentences). A second manipulation involved the nature of the sentences in the training: syntactically and semantically preserved (but vocoded) English sentences, versus semantically empty "Jabberwocky" sentences, versus strings of nonwords. The authors found that listeners' perception of vocoded speech did not improve in the nonword condition. They thus concluded that although learning occurs at the phonological level (since improved performance extended to words that were not encountered in the training set), it is mediated by lexical representations.

However, a recent follow-up study modified this conclusion (Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008). Hervais-Adelman et al. hypothesized that listeners in their previous study might have failed to learn from sentences made up of nonwords because such stimuli could be quite difficult to keep in memory, preventing perceptual learning processes from operating on them. They tested this idea by using a series of individual words, or a series of individual nonwords, as the training stimuli. Because each item was at most two syllables long, listeners could encode the nonwords about as well as the words. Under these conditions, essentially the same amount of perceptual learning occurred with nonword training items as with real word training items. Thus, what seems to be critical is that the degraded speech must be encodable into some kind of working memory representation. When that can be done, then adjustments can be made at the phonological level.

A major motivation for studying vocoded speech is that it shares important properties with the signal that is provided by cochlear implants; in both cases, the auditory system is fed information from a small number of frequency-based channels, rather than from the continuous frequency range in normal hearing. Several authors have taken this one step further, and added a frequency shift to the vocoded speech. This is comparable to an additional complication inherent in cochlear implants: Because of surgical limitations, the implanted electrodes do not span the normal frequency range—they are skewed upward. Rosen, Faulkner, and Wilkinson (1999) therefore both vocoded and frequency-shifted speech that they played

to normal-hearing listeners. Initially, the listeners could recognize only about 1% of the words in these vocoded/shifted sentences. However, after nine 20-min training sessions with these materials, performance increased to approximately 30% correct. Fu and Galvin (2003), using similar materials, also found that recognition improved with training. Most recently, Stacey and Summerfield (2007) reported that with four sessions of 1–2 h, there was a 13%–18% improvement in word recognition. As in the work of Bradlow et al. (1997; Bradlow et al., 1999), these authors found that generalization performance was improved if the training was done with multiple talkers, rather than with a single speaker.

As with compression and vocoding, synthesized speech can initially be quite difficult to understand. Nusbaum and his colleagues (Fenn, Nusbaum, & Margoliash, 2003; Greenspan, Nusbaum, & Pisoni, 1988) examined perceptual learning by presenting listeners with speech generated by a relatively crude speech synthesizer and looking for improved transcription performance as a function of experience with the synthetic speech. Fenn et al. focused on the interesting question of whether such learning gets consolidated during sleep, as other (nonlinguistic) perceptual learning appears to. Listeners were trained by hearing 150 synthesized monosyllables with the printed version of each word provided as feedback. They responded by typing the words, and then went through the same words without feedback. Learning was measured by presenting a different set of 100 words for transcription, without feedback. With this rather minimal training, listeners' comprehension of the synthetic speech nonetheless improved by about 20% on tests immediately after training. When they were tested 12 h later, their performance varied as a function of whether they had slept in the intervening time period or not. Fenn et al. ran a large number of conditions to tease apart sleep from the amount of time between training and test, the time of day of training, or the time of day of testing. They found a clear pattern: Subjects who were tested 12 h after training with no intervening period of sleep declined from 20% improvement down to about 10%. However, subjects who were tested 12 h later *with* an intervening sleep period showed no such decline: Their improvement was essentially the same as the original learning advantage. Thus, this study indicates that perceptual learning for speech, like perceptual and motor learning in other nonlinguistic domains, may require sleep to consolidate learning.

THEME II
Perceptual Learning As Phonetic Retuning

We turn now to a relatively newer area of inquiry, termed "perceptual learning" by some (e.g., Norris, McQueen, & Cutler, 2003), and "recalibration" by others (Bertelson, Vroomen, & de Gelder, 2003). The studies in this area are quite diverse, addressing a range of questions and employing various types of stimuli and procedures. What unifies this area of inquiry is a focus on phonetic categories and how they are retuned to become more aligned with the input. This theoretical focus naturally leads to questions about *how* such retuning occurs and about what the effects of such retuning are for subsequent processing of speech.

The literature on phonetic retuning addresses many of the same issues that have come up in the literature we have reviewed in Theme I. These studies have examined what information is required for perceptual learning, how quickly such learning can occur, and how long it lasts. There are also investigations of when and how the retuning may go away, the level of specificity of the retuning (both with respect to the particular speakers and to the particular phonemes encountered at training), and the kinds of information that might block retuning.

Although the details vary from study to study, the basic procedure is to present listeners with phonetically ambiguous stimuli, with some source of contextual information that disambiguates the stimuli. Perceptual learning is subsequently indexed by a shift in phonetic categorization toward the contextually defined speech environment. The measurement is one of category boundary location, assessed by having the listeners identify members of a continuum of speech sounds. After exposure to acoustically ambiguous speech sounds that are contextually disambiguated, listeners increase their report of sounds consistent with context they received. Presumably such shifts should help the listener understand speech better in the prevailing input environment. By focusing specifically on phonetic boundary shifts, these studies provide a clear indication of exactly what is changing in perceptual processing as a function of experience. An important advantage of such a targeted approach is that these studies provide insight as to *how* phonetic space is remapped, something that is generally not available in the studies we reviewed in Theme I.

The study of perceptual learning as phonetic retuning began with two seminal articles, both published in 2003 (Bertelson et al., 2003; Norris et al., 2003). The two articles took different approaches but came to the same conclusion: Contextual knowledge (specifically, lexical information in Norris et al., 2003, and visual information in Bertelson et al., 2003) guides perceptual learning. We will examine studies based on Norris et al.'s lexical manipulation first, followed by studies based on Bertelson et al.'s audiovisual approach.

II.1. Lexically Induced Perceptual Learning

Norris et al. (2003) presented Dutch listeners with the speech of a single female Dutch talker. The talker's speech was manipulated so that she seemed to produce instances of a particular fricative (/s/ for one group of listeners, /f/ for another) in an ambiguous way (i.e., as a sound midway between [f] and [s]; hereafter, [?]). The sound was presented to listeners in one of three conditions: an [f]-training condition, in which the ambiguous sound [?] replaced [f] in words such as *witlof* (meaning "chicory"; note that *witlos* is not a Dutch word); an [s]-training condition, in which [?] replaced [s] in words such as *naadlbos* ("pine forest"; again, *naaldbof* is not a Dutch word); or at the end of nonwords, where neither a final [f] or [s] would create a word in Dutch. Norris et al. predicted that

if listeners use lexical knowledge to guide their interpretation of the ambiguous fricative, listeners who heard [?] in [f]-final words would subsequently categorize more sounds on an [ɛs]–[ɛf] continuum as [f], whereas those who heard the same sound in [s]-final words would categorize more items as [s]. Listeners who heard [?] at the end of nonwords should show no bias toward [f] or [s]. The results confirmed these predictions, and suggest that listeners are able to use lexical knowledge to guide their interpretation of acoustic–phonetic information.

In subsequent studies (McQueen, Cutler, & Norris, 2006; McQueen, Norris, & Cutler, 2006), the authors replicated the original Norris et al. (2003) finding using slightly different tasks at test and at exposure, respectively. McQueen, Cutler, and Norris used the same (lexical decision) task at exposure as Norris et al., but changed the test phase in order to investigate whether representations are abstract or perceptually detailed episodes. McQueen, Norris, and Cutler looked at whether the perceptual learning effect is automatic, or whether it might depend on explicit decisions about the lexical status of the items containing the ambiguous pronunciations. The results of both studies confirmed the original perceptual learning results and additionally suggested that (1) perceptual learning generalizes to words outside the original training set (McQueen, Cutler, & Norris, 2006) and that (2) the learning effect arises automatically as a consequence of simply hearing these ambiguous pronunciations in words (McQueen, Norris, & Cutler, 2006). Cutler, McQueen, Butterfield, and Norris (2008) expanded this conclusion to include a role for phonotactic information, by demonstrating that listeners could also make use of prelexical constraints to guide perceptual learning. Thus it seems that lexical access itself is not required, but that some constraint on how the sound should be interpreted is necessary.

Although lexical constraint is not necessary to drive perceptual learning, it is a powerful source of context. Leach and Samuel (2007) measured the growth of perceptual learning shifts as a function of the establishment of lexical representations. Over the course of 5 days, listeners were taught a set of new "words" (e.g., *nomemsolly*). In addition to many presentations of these new words with all of their sounds correctly pronounced, each day a small number of presentations included an ambiguously produced fricative (midway between [s] and [ʃ]). As the new words became lexicalized, these novel pronunciations were increasingly able to generate perceptual learning. Leach and Samuel found that the growth of perceptual learning depended on how the listeners learned the new words: When words were learned by associating them with pictures of unusual objects, the words developed a strong ability to generate perceptual learning, whereas when the words were learned in the context of making phoneme present/absent judgments, they were not able to generate strong perceptual learning.

Given that recalibration of the phonetic boundary is driven by context, it is natural to ask whether the shift reflects a true perceptual change or a postperceptual decision bias. Clarke-Davidson, Luce, and Sawusch (2008) recently addressed this question, using a discrimination task and a signal detection analysis. Both approaches suggested that perceptual learning is based on a true representational shift (i.e., a remapping of acoustic–phonetic space), rather than on a shift in decision criteria.

If perceptual learning does indeed involve a remapping of acoustic–phonetic space, then once phonetic categories have changed, how and when do they return to "normal"? Two studies suggest that learning is quite long-lived and resistant to change. Kraljic and Samuel (2005) demonstrated that perceptual learning to a noncanonical pronunciation of [s] or [ʃ] remained robust after a 25-min delay. Moreover, it remained robust even after listeners heard many canonical pronunciations of [s] and [ʃ] during the 25-min delay. The only condition in which the perceptual learning effect attenuated at all was when listeners heard canonical pronunciations of [s] and [ʃ] *from the same speaker* that they had originally adjusted to. Eisner and McQueen (2006) subsequently showed that learning remains stable over a much longer delay—12 h—regardless of whether subjects slept in the intervening 12 h (their "night group") or not (their "day group"). Unlike in Fenn et al.'s (2003) study of perceptual learning for synthetic speech, Eisner and McQueen (2006) did not find evidence for consolidation effects. The shifts after sleeping were no larger than those without sleep; they neither grew nor decayed.

The lack of any decay in perceptual learning, which suggests an enduring change to the representations, potentially causes complications if these shifts are not speaker specific. Otherwise, how could a new speaker's pronunciations be learned once the representations have been tuned to reflect those of a previous speaker? Recall that Kraljic and Samuel (2005) found that canonical pronunciations from the same speaker, but not from a different speaker, attenuated perceptual learning (slightly), consistent with speaker specificity.

In a direct test of this issue, Eisner and McQueen (2005) also found evidence that learning is speaker specific. The authors first replicated Norris et al.'s (2003) results, showing perceptual learning of a female speaker's pronunciation when listeners heard words with the ambiguous mixture of [s] and [f] and were then tested on an [ɛs]–[ɛf] continuum that was made entirely from her voice. The critical new finding was that perceptual learning was also found when the continuum vowel ([ɛ]) was spoken by a novel talker (both a novel male and a novel female), as long as the fricatives were from the original talker's speech. When the continuum was created entirely from the speech of a novel talker, there was no perceptual learning—unless, unbeknownst (and undetectable) to the listener, the novel talker's fricatives had been spliced into the original talker's speech during exposure.

The speaker-specific results reported by Eisner and McQueen (2005) and by Kraljic and Samuel (2005) were based on studies using fricatives as the critical phonemes. Kraljic and Samuel (2006) investigated perceptual learning for stop consonants, and obtained different results. They exposed listeners to either a male or a female talker with

an ambiguous stop, one midway between [d] and [t]. Listeners were tested both on stimuli from the same speaker they heard during exposure, and on stimuli from a very different speaker. Unlike the results from the studies using fricatives, listeners not only showed significant perceptual learning effects for the same speaker, they showed equally robust effects for the new speaker. Interestingly, learning also generalized to new *phonemes* that shared the same voicing feature distinction as [d] and [t]—specifically, listeners generalized the feature they had learned for [d] and [t] to the (previously unheard) stop consonants [b] and [p].

Kraljic and Samuel (2007) addressed speaker specificity from a slightly different angle. They investigated whether listeners could simultaneously show learning for the pronunciations of more than one talker. Listeners heard blocks of words (containing critical ambiguous phonemes) from two different voices, with opposite directions of potential retuning (e.g., one voice had [?] in words supporting a [d] interpretation, whereas the other voice had [?] in [t] contexts). The results were consistent with the prior studies: The perceptual learning was speaker specific when the critical phonemes were the fricatives [s] and [ʃ], but when the critical phonemes were stop consonants ([t] and [d]), listeners did not make speaker-specific adjustments. Rather, they appeared to adjust for one speaker's pronunciation, and then subsequently readjust the representation to be in line with the new speaker's different pronunciation. The result was that for the stops, the perceptual learning reflected the most recent pronunciation heard, regardless of the speaker.

The ability of listeners to rapidly adjust their representations to be congruent with the input they get raises a question that is at the heart of the field of speech perception: How does the perceptual system balance such flexibility with the stability that is necessary for the reliable perception of phonemes? Kraljic, Samuel, and Brennan (2008) suggested that one answer might be that learning is somehow constrained to those pronunciations that are likely to remain stable. They found that perceptual learning is subject to a primacy bias: Pronunciations that are heard upon initial exposure to a speaker are learned, whereas those same pronunciations are not learned if they do not form part of the initial listening experience. Kraljic, Samuel, and Brennan (2008) also found that listeners did not learn a pronunciation if it could be attributed to some transient alternative (speaker-external) factor, such as a pen in the speaker's mouth.

A second line of research provides converging evidence that perceptual learning is constrained in some contexts. Kraljic, Brennan, and Samuel (2008) exposed listeners either to a speaker whose ambiguous pronunciation of [s] occurred intervocalically (as a contextually independent ambiguity) or to a speaker whose ambiguous [s] occurred only in the context of a following [tr]. This context was chosen because in many American English dialects, the [s] in this position is naturally produced as a sound that is ambiguous, somewhere between [s] and [ʃ]. Despite the identical acoustics in this contextually constrained case and the contextually independent one, perceptual learning occurred only in the latter.

The perception of nonambiguous sounds. The perceptual learning studies we have discussed so far investigate the effect that exposure to a pronunciation has on subsequent processing of that pronunciation. We have been discussing perceptual learning as a shift in representational space. An alternative possibility, however, is that learning reflects the tuning of transformational processes that enable the listener to more efficiently translate ambiguous input to the (unshifted) representation. Dahan, Drucker, and Scarborough (2008) addressed this question by looking at the effect that exposure to one nonstandard pronunciation (e.g., raised [ae] before /g/, as in some speakers' pronunciations of the word *bag*) has on subsequent perception of a different, related pronunciation (e.g., standard [ae] before /k/, as in the word *back*). Using eyetracking as their measure, Dahan et al. found that exposure to a speaker with a raised vowel (in words such as *bag*) caused listeners to be more likely to interpret the standard pronunciation /bae/ as the onset of /back/ (as opposed to finding the onset ambiguous between /bag/ and /back/, as they would before exposure). This finding is novel because it goes beyond demonstrating that listeners can learn to adapt to a nonstandard pronunciation (all the work we have so far reviewed has shown that they can). Instead, it suggests that as listeners learn, they adjust their probabilities for particular sounds to be mapped onto particular words, and thereby affect interpretation of a standard pronunciation that does not require adaptation (ruling out an account of learning that is based on transformation of the signal).

Production. Kraljic, Brennan, and Samuel (2008) examined the effect of perceptual learning on subsequent production. Subjects produced words to complete a story before and after a perceptual learning manipulation. Although the perceptual system showed robust perceptual learning, the production system did not make a corresponding change: Perceptual exposure did not result in changes in production. However, as Kraljic, Brennan, and Samuel discussed, there are a number of reasons (many related to the experimental procedure) why speakers might not have changed their production. Thus, the question of how perceptual learning affects production is still very much open.

Remapping of vowel space. Maye, Aslin, and Tanenhaus (2008) looked at the remapping of a vowel category. They showed that listeners adapted to a dialectal variant such that a stimulus that they had originally perceived as a nonword (e.g., *wetch*) came to be perceived as a word (e.g., as the word *witch*). Maye et al. also found that learning reflected very targeted perceptual shifts in the direction experienced in the dialect (i.e., toward lower front vowels)—it did not simply cause listeners to relax the boundaries of what constitutes a particular vowel. On the other hand, Maye et al.'s results also suggest that subjects do not completely remap their vowel space, because listeners accepted both accented and standard English pronunciations of words after perceptual learning (i.e., they

accepted both "wetch" and "witch" as versions of the word "witch").

II.2. Audiovisually Induced Perceptual Learning

As we noted at the outset, studies of perceptual learning that focus on phonetic boundary changes began with a pair of seminal studies, one of which used lexical context (Norris et al., 2003), and one of which used visual cues as the context (Bertelson et al., 2003). We turn now to the studies based on the latter.

Bertelson et al. (2003) began the investigation of auditory perceptual learning (in their term, *recalibration*) guided by audiovisual integration. In their methodology, listeners heard three different acoustic tokens: [aba], [ada], or [a?a], the latter being a token that was ambiguous between /b/ and /d/ for that listener. Each token was dubbed onto a video of a speaker articulating /aba/ or /ada/. Identification tests during this exposure phase showed that the visual context produces a very strong immediate bias: Listeners heard the ambiguous token as whatever the face they saw was articulating. A subsequent auditory-only test demonstrated that this visually induced bias results in subsequent perceptual learning: Tokens that were formerly ambiguous were heard as /b/ if the subject had previously seen such tokens with a face articulating /b/, but were heard as /d/ if the visual exposure had been a /d/.

Bertelson et al. (2003) included a control condition in which the auditory part of the initial audiovisual exposure was an unambiguous token (either /b/ or /d/), rather than the token chosen to be ambiguous for the listener. Previous research (e.g., Eimas & Corbit, 1973; Samuel, 1986) has shown that repeated exposure to an unambiguous sound reduces report of sounds similar to the repeating one (e.g., hearing /ba/ many times reduces subsequent report of /b/ for a /ba/–/da/ test continuum). Bertelson et al. replicated this result. Critically, the shifts with unambiguous repeating sounds ("selective adaptation") are in the opposite direction from those produced by perceptual learning, making it clear that these are different phenomena.

Vroomen, van Linden, Keetels, de Gelder, and Bertelson (2004) followed up on the differences between selective adaptation and recalibration. Specifically, they looked at the time course with which each of these effects would dissipate. Contrary to what we have seen for lexically induced perceptual learning of fricatives (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), visually induced perceptual learning dissipated very quickly: It lasted only over the first 6 test trials. For test trials 7–20, the effect instead looked like a selective adaptation effect: Listeners began to categorize fewer items as being like the biased phoneme, rather than more of them.

In a follow-up study, Vroomen, van Linden, de Gelder, and Bertelson (2007) replicated these results and also examined the build-up of perceptual learning and selective adaptation. They inserted test trials after 1, 2, 4, 8, 32, 64, 128, or 256 exposure tokens. They found that perceptual learning occurs very rapidly: Listeners demonstrated learning after a single exposure token. This learning increased through about 8 exposures, after which it reached a plateau and then began to decrease. These results indicate that perceptual learning effects produced by audiovisual biases dissipate rather quickly, as Vroomen et al.'s (2004) study had suggested. In contrast, selective adaptation took a little longer to materialize, but continued to grow throughout the entire range of exposure that was tested.

Vroomen and Baart (2009a) have used the contrast between selective adaptation and recalibration to demonstrate that the former is a more automatic, acoustically driven phenomenon than the latter. The key to their test was their use of sine-wave-speech stimuli (Remez, Rubin, Pisoni, & Carrell, 1981). Sine wave speech is made by tracking the formant paths of speech, and reducing the formants to time-varying sine-wave patterns; typically, just the first three formants are used, with each of these formants replaced by a sine wave that follows the center frequency of the formant. Most listeners initially hear such stimuli as a kind of varying whistling sound. However, if instructed that this is speech, many listeners are able to hear it as such, and to understand the words being spoken. Vroomen and Baart (2009a) tested two groups of subjects with sine wave versions of a continuum that ranged between /omso/ and /onso/. One group was told to label each token as either "omso" or "onso"; the other group used the labels "1" and "2." The first group thus was encouraged to hear these items as speech, whereas the second group was not. Following this experience, adaptation and recalibration tests were run using the typical procedures: For adaptation, endpoint /omso/ and /onso/ tokens were paired with matching videos, and these congruent stimuli were presented repeatedly; subjects labeled (auditory only) members of the test series between rounds of adaptation. For recalibration, an ambiguous member of the continuum was paired with a video that was either clearly /omso/, or one that was clearly /onso/. The results were very clear. On the adaptation test, the usual contrastive shifts were observed both for the group that had learned to hear the stimuli as speech, and for the group that heard them as whistles. For the recalibration test, only the group that heard the stimuli as speech showed recalibration—subjects in the nonspeech group did not shift their boundaries at all. These results indicate that adaptation is being driven by the sounds, regardless of their perceptual interpretation, whereas recalibration occurs only when there is a conflict between auditory speech and visual speech.

Given the results that we have reviewed above, it would be tempting to conclude that visually induced perceptual learning results in a transient shift, whereas lexically induced learning, for some as yet unknown reason, results in a much more stable change. But there are many other differences between the types of studies looking at lexically induced learning and those looking at visually induced learning. One critical factor may be the nature of the critical phonemes. Recall that the long-lasting lexically driven perceptual learning effects were obtained with fricatives; all of the audiovisual experiments have used stops as the critical phonemes (and, as Kraljic & Samuel's [2007] perceptual learning study showed, stops result in quite different learning effects than fricatives). Van Linden and

Vroomen (2007) directly compared visually and lexically induced learning, using the same experimental procedure and the same test stimuli (stop consonants [t] and [p]). They found that exposure to audiovisual and to lexical stimuli resulted in statistically equivalent learning, and that the learning effects for both dissipated quickly, at similar rates. Both kinds of learning increased when the exposure set included not only the ambiguous token, but also the contrasting unambiguous token. However, even when larger, the effects still dissipated quickly; van Linden and Vroomen (2007) suggested that dissipation is the result of a shifting criterion during prolonged testing. Their study included one other methodological comparison of the lexical and audiovisual literatures: Exposing listeners to one and then the other pronunciation (the standard method in audiovisual studies) produced the same learning effect as did exposing them to only one pronunciation and comparing them with listeners who had learned the other pronunciation (the standard method in lexical studies).

Recall that Eisner and McQueen (2006) tested whether lexically driven perceptual learning (for fricatives) changes over substantial delays (12 h, either during the same day, or overnight), and that they found neither a dissipation nor any evidence for consolidation effects. Vroomen and Baart (2009b) examined whether, when immediate visually induced learning was robust, it might also return after some consolidation had been able to occur—specifically, after a 24-h period. They found that despite large immediate effects, learning dissipated quickly and did not return after a 24-h delay. This comparison may be qualified by the fact that Eisner and McQueen (2006) tested fricatives in their lexically driven perceptual learning study, whereas Vroomen and Baart (2009b) used stops.

Although the findings of Vroomen et al. (2004; Vroomen et al., 2007) and of Vroomen and Baart (2009b) for audiovisual recalibration are in contrast to what has been found with lexically induced learning of fricatives, the direct comparison of the two kinds of learning (van Linden & Vroomen, 2007) found more similarities than differences. When the experimental procedure and stimuli were held constant, they resulted in similar learning, dissipation at similar rates, with similar effects of the same kinds of additional information. The only consistent difference seems to be that visually induced learning produces numerically larger effects than lexically induced learning (see van Linden & Vroomen, 2007, for a discussion). At this point, there is no clear understanding for the ways in which lexical and visual context differ in producing perceptual learning. It is conceivable that the different information sources might drive different types of learning. Recall that Goldstone (1998) described four different mechanisms of perceptual learning: A new "detector" can be developed, there can be improved differentiation of cues, unitization of cues can improve, or different attentional weights can develop for the available cues. It is possible that lexical context facilitates one of these four mechanisms, and that visual speech supports a different one, with the observed differences following from characteristics of the mechanisms.

Recently, van Linden and Vroomen (2008) explored whether there is a developmental trend in the ability to use visual information to learn aspects of a speaker's pronunciation. They tested 5-year-olds and 8-year-olds on the visually induced learning paradigm, and found that the 8-year-olds showed significant learning, but the 5-year-olds did not. Thus, there is evidence for a developmental trend for learning to categorize a sound in accord with previously experienced lipread information.

As van Linden and Vroomen (2008) pointed out, there are many possible interpretations of these results. In fact, these findings must be considered within the context of a recent study by Teinonen, Aslin, Alku, and Csibra (2008). These authors examined audiovisual learning effects in 6-month-olds. The infants were exposed to a distribution of synthetic speech sounds from a /ba/–/da/ continuum. Critically, the distribution was "unimodal"—most of the tokens were from near the phonetic boundary, with decreasing numbers as the stimuli became more clearly /ba/ or more clearly /da/. Two groups were formed, based on the video that accompanied these syllables. In one group, a face articulating /ba/ accompanied the four tokens from the /ba/ side of the continuum, and a face articulating /da/ was paired with the four tokens from the /da/ side. In the other group, a single video was paired with all eight speech tokens (for half of the infants it was the /ba/ video, and for the other half it was the /da/ video). After this exposure phase, the authors presented the speech syllables without any video, and tested for categorization. They found that infants who had experienced the differential video information tended to divide the speech into two categories, whereas the other infants did not. These results show that the visual information during exposure influenced the later categorization of speech, a result that is essentially the same as in the recalibration studies we have considered.

CONCLUSION

As we noted at the outset, researchers have known since at least the time of Volkmann (1858) that the perceptual system can fine-tune its performance through experience. Over the years, investigators have provided considerable additional evidence for such perceptual learning, consistent with Goldstone's (1998) definition: "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment" (p. 586). In this article, we have reviewed several literatures that have developed in the study of perceptual learning for speech. Within Theme I, we looked at three approaches that clearly fit within Goldstone's definition: cases in which people learn to discriminate foreign language contrasts, cases in which people adapt to unusual dialects or accents, and cases in which people improve in their ability to understand degraded speech. In all of these cases, the measurements assess improvements in the listeners' ability to respond to the environmental input.

For the two literatures we considered in Theme II, such an improvement in processing speech is not directly given

in the measurements, because what is measured is a phonetic boundary shift. However, it is quite plausible, even likely, that these shifts do indeed improve the listeners' ability to perceive the prevailing environmental input, since the shifts reflect that input. Moreover, the studies in Theme II provide very specific information about how perceptual learning can be implemented in the system that must deal with the complex and noisy signal that speech is: The system recalibrates its categorization boundaries, using whatever contextual information is reliably provided in the environment. In this sense, it is quite fortuitous that two rather different approaches to this problem happened to arrive at the same time—Norris et al.'s (2003) work using lexical context, and Bertelson et al.'s (2003) study of audiovisual recalibration. The similarities in these two literatures suggest that the perceptual system is aggressively opportunistic in its adaptation to the environment in which it must operate. What is truly remarkable is that our perceptual systems seem to have evolved in such a way that they can make these kinds of significant adjustments, while at the same time providing us with the impression of a stable and consistent perceptual world.

AUTHOR NOTE

Preparation of this article was supported by NIMH Grant R01-051663, NSF Grant 0325188, and NIH Postdoctoral Training Grant F32 HD052342. We thank James Sawusch and an anonymous reviewer for their helpful suggestions. Correspondence concerning this article should be addressed to A. G. Samuel, Department of Psychology, State University of New York at Stony Brook, Stony Brook, NY 11794-2500 (e-mail: asamuel@ms.cc.sunysb.edu).

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.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, **32**, 233-250.

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

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation hypothesis. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception* (pp. 167-224). Cambridge, MA: MIT Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America*, **109**, 775-794.

Bradlow, A. R., Akahane-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**, 977-985.

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

Bradlow, A. R., Pisoni, D. B., Akahane-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**, 2299-2310.

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

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, **70**, 604-618.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. *Proceedings of Interspeech 2008*, Brisbane, Australia.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, **34**, 269-284.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, **108**, 710-718.

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

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

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, **4**, 99-109.

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

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

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, **425**, 614-616.

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. In C. Ferguson and T. Huebner (Eds.), *Crosscurrents in second language acquisition and linguistic theories*. Philadelphia: John Benjamins.

Flege, J. E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, **16**, 425-442.

Fu, Q.-J., & Galvin, J. J., III (2003). The effects of short-term training for spectrally mismatched noise-band speech. *Journal of the Acoustical Society of America*, **113**, 1065-1072.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, **62**, 32-41.

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

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, **49**, 585-612.

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

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

Kingston, J. (2003). Learning foreign vowels. *Language & Speech*, **46**, 295-349.

Klatt, D. H. (1976). The linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, **107**, 54-81.

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

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

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

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, **19**, 332-338.

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

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese

listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, **94**, 1242-1255.

LIVELY, S. E., PISONI, D. B., YAMADA, R. A., TOHKURA, Y., & YAMADA, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, **96**, 2076-2087.

LOGAN, J. S., LIVELY, S. E., & PISONI, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

MAYE, J., ASLIN, R. N., & TANENHAUS, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, **32**, 543-562.

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

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

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

NORRIS, D., MCQUEEN, J. M., & CUTLER, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, **47**, 204-238.

NYGAARD, L. C., & PISONI, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, **60**, 355-376.

NYGAARD, L. C., SOMMERS, M. S., & PISONI, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.

PALLIER, C., SEBASTIÁN-GALLÉS, N., DUPOUX, E., CHRISTOPHE, A., & MEHLER, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition*, **26**, 844-851.

PETERSON, G. E., & LEHISTE, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America*, **32**, 693-703.

REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

ROSEN, S., FAULKNER, A., & WILKINSON, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America*, **106**, 3629-3636.

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

SEBASTIÁN-GALLÉS, N., DUPOUX, E., COSTA, A., & MEHLER, J. (2000). Adaptation to time-compressed speech: Phonological determinants. *Perception & Psychophysics*, **62**, 834-842.

STACEY, P. C., & SUMMERFIELD, A. Q. (2007). Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *Journal of the Acoustical Society of America*, **121**, 2923-2935.

STRANGE, W., JENKINS, J. J., & JOHNSON, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, **74**, 695-705.

TEINONEN, T., ASLIN, R. N., ALKU, P., & CSIBRA, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, **108**, 850-855.

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

VAN LINDEN, S., & VROOMEN, J. (2008). Audiovisual speech recalibration in children. *Journal of Child Language*, **35**, 809-822.

VOLKMANN, A. (1858). Ueber den Einfluss der Uebung auf das Erkennen räumlicher Distanzen. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Classe*, **10**, 38-69.

VROOMEN, J., & BAART, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition*, **110**, 254-259.

VROOMEN, J., & BAART, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language & Speech*, **52**, 341-350.

VROOMEN, J., VAN LINDEN, S., DE GELDER, B., & BERTELSON, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, **45**, 572-577.

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**, 55-61.

WANG, Y., SPENCE, M. M., JONGMAN, A., & SERENO, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, **106**, 3649-3658.

(Manuscript received November 5, 2008;
revision accepted for publication March 26, 2009.)