Ecological Psychology, 26:4–15, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1040-7413 print/1532-6969 online
DOI: 10.1080/10407413.2014.874839

Routledge
Taylor & Francis Group

Phonetic Category Learning and Its Influence on Speech Production

Richard N. Aslin
Department of Brain and Cognitive Sciences
University of Rochester

One of the hallmarks of any flexible system of perception and motor control is the ability to adjust to changes induced by dialect, development, fatigue, disease, or aging. Phonetic categories are an essential component of language that enables listeners and speakers to communicate effectively. Four studies are reviewed that illustrate how adults and infants adjust their phonetic categories rapidly and efficiently to maintain a tight coupling between speech perception and speech production. Although this process of adaptive plasticity takes place at the level of phonetic categories, it is also constrained by the lexicon. Words that share similar sounds or similar vocal-articulatory gestures impede the process of adaptation.

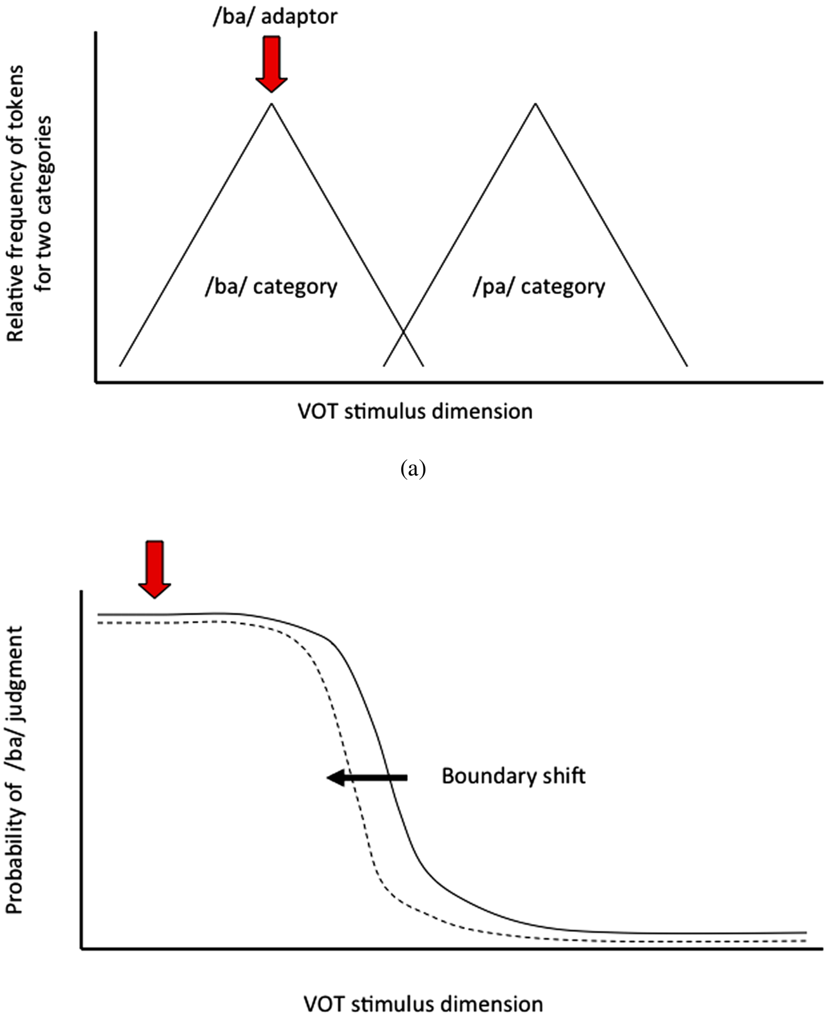
The study of speech perception has a long and interesting history (Raphael, Borden, & Harris, 2007). An early view was that mature listeners process speech in a manner quite distinct from basic psychoacoustic principles, in part because speech perception appears to violate fundamental tenets of nonspeech processing such as Weber's Law and a variety of Gestalt principles (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This *Speech-Is-Special* viewpoint was supported by classic findings on categorical perception (CP) of stop consonants such as /b/ and /p/—labeling of these consonants was perfectly predicted by the ability to discriminate small differences in an acoustic parameter called voice onset time (VOT). The same physical difference in VOT that was easily discriminated when it straddled the boundary between two categories (thus readily labeled as /b/ or /p/) elicited chance discrimination performance when

Correspondence should be addressed to Richard N. Aslin, Department of Brain and Cognitive Sciences, Meliora Hall, River Campus, University of Rochester, Rochester, NY 14627. E-mail: aslin@cvs.rochester.edu

both VOT tokens came from the same category (either both /b/ or both /p/). Despite the fact that the canonical view of CP was known to be incorrect in the mid-1970s (Pisoni & Tash, 1974), it has persisted as settled dogma for over 40 years (see further evidence of within-category sensitivity in adults by McMurray, Tanenhaus, & Aslin, 2002, and in infants by McMurray & Aslin, 2005). Moreover, the notion that speech-is-special—although certainly true in a general sense—is not supported by specific corollaries of that theory. Notably, (a) some *nonspeech* sounds are also perceived categorically, even by infants; (b) infants perceive many *nonnative* speech sounds categorically, despite never having been exposed to these sounds; and (c) several *nonhuman* species perceive speech categorically. Why does the canonical view of CP persist in the face of this countervailing evidence? One reason is that CP captures a fact about the close coupling of speech production and speech perception. Speech sounds produced by a given talker form a distribution of tokens that lie along one or more acoustic dimensions, and listeners interpret those tokens in such a way that they map seamlessly onto the intended target (e.g., the /b/ or /p/ category) of that talker's productions.

In the past decade, there has been a resurgence of interest in the mechanisms of adaptive plasticity that enable speech perception and speech production to maintain the accuracy of communicative intent between speaker and listener. One important observation made in recent years—even though recognized in qualitative form decades ago—is that listeners must be sensitive to the distributional properties of the speech sounds to which they are exposed in their listening environment. That is, in addition to *talker-specific* distributions, there are also *dialect* or *talker-general* distributions that represent the aggregate of the acoustic/phonetic variations to which listeners in a given native-language dialect are exposed (see Figure 1a). Speakers must produce speech tokens that fall within these distributions or risk being misinterpreted, and listeners must assign these speech tokens to the appropriate phonological category to avoid misunderstandings.

There is compelling evidence from several decades of research on infant speech perception (Kuhl, 2004; Werker, Yeung, & Yoshida, 2012) that these phonetic categories are tuned by early exposure to massive amounts of distributional information. Interestingly, this exposure *narrows* initially exuberant discriminative sensitivity. That is, in contrast to most findings from studies of development, infants are *more* sensitive than their parents to phonetic distinctions that are *absent* in their native language. Werker and Tees (1984) showed that 6-month-olds from an English-speaking environment could discriminate a non-English phonetic contrast that their parents could not discriminate. But only a few months later these same infants were unable to make this same phonetic discrimination, becoming adultlike presumably by implicitly learning that certain distributions were not attested in their language environment. Maye, Werker, and



(a)

(b)

FIGURE 1 (a) Schematic of two phonetic categories and the repetition of a single category adaptor. (b) Result of selective adaptation on category labeling (color figure available online).

Gerken (2002) and Maye, Weiss, and Aslin (2008) showed that similar changes in phonetic discrimination could be induced by short-term laboratory exposure.

Adults are not immune to being susceptible to changes in these distributional properties of acoustic/phonetic information. In classic experiments under the rubric of selective adaptation, it was shown that adults' judgments of phonetic-category membership were influenced by listening to a single, repeated speech token. When this token came from the peak of the phonetic category distribution (i.e., the prototype), subsequent judgments of tokens near the category boundary were less likely to be assigned to that prototype (see Figure 1b). That is, a narrowing of the distribution around the mean of the category—in the most extreme way by only presenting tokens at the mean, with zero variance—had the effect of restricting the interpretation of exemplars that deviated from that prototype. If a postadaptation test token was too far from the category mean, it was now more likely to be interpreted as a member of the adjacent category along the phonetic dimension being manipulated.

In the sections that follow, four studies conducted over the past few years by my students and colleagues are reviewed. Each bears on the foregoing question of how phonetic categories undergo reorganization in response to changes in the distributional properties of speech input. The bottom line from these studies is that adults and infants are remarkably sensitive to these distributional properties and use them to fine-tune their interpretations of speech signals in an ongoing and rapid manner. Moreover, these adaptive processes occur both at the level of syllables and at the level of words, suggesting that the manner in which the lexicon—the mental dictionary—is organized plays an important role in this adaptation process. Finally, sensitivity to distributional information in speech categories plays itself out in the control of motor commands to the vocal-articulatory apparatus so that speech production, as well as speech perception, is being updated continuously to adapt to fatigue, development, and other perturbations of the production system.

DISTRIBUTIONAL LEARNING OF SPEECH CATEGORIES

As summarized earlier, the selective-adaptation paradigm provided a stark contrast between the natural variability of speech productions to which a listener is exposed by narrowing that distribution to a single token at the category mean. Meghan Clayards, in her dissertation (see Clayards, Tanenhaus, Aslin, & Jacobs, 2008), asked a more subtle question: Do listeners update their phonetic category judgments based on the *variance* of tokens drawn from the category? As shown in Figure 2, the *means* of the two categories lying along a VOT continuum were not altered, but the variances of the two category distributions differed

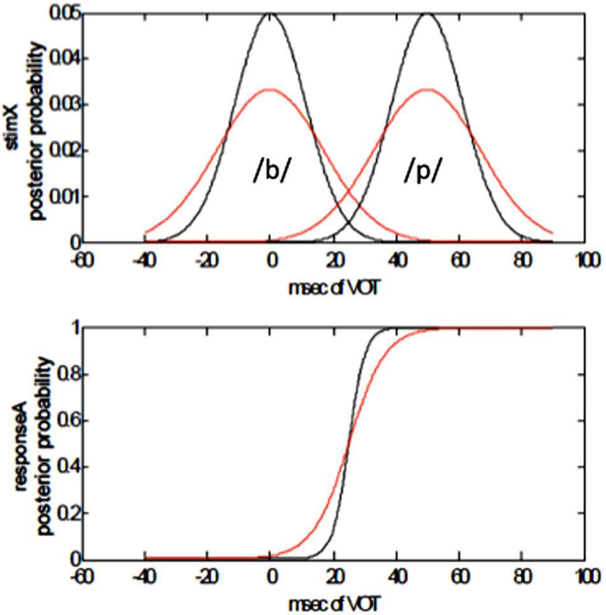


FIGURE 2 Narrow (black) and wide (red) distributions and predicted labeling functions (color figure available online).

across exposure conditions for two groups of adults. The means were chosen to match the existing prototypes of the dialect from which the participants were sampled, and the variance of the *wide* distribution was chosen to match the existing variance of this dialect. At issue was whether adults assigned to the *narrow* distribution would implicitly learn that the sharpness of the category boundary was steeper compared with the wide (control) condition.

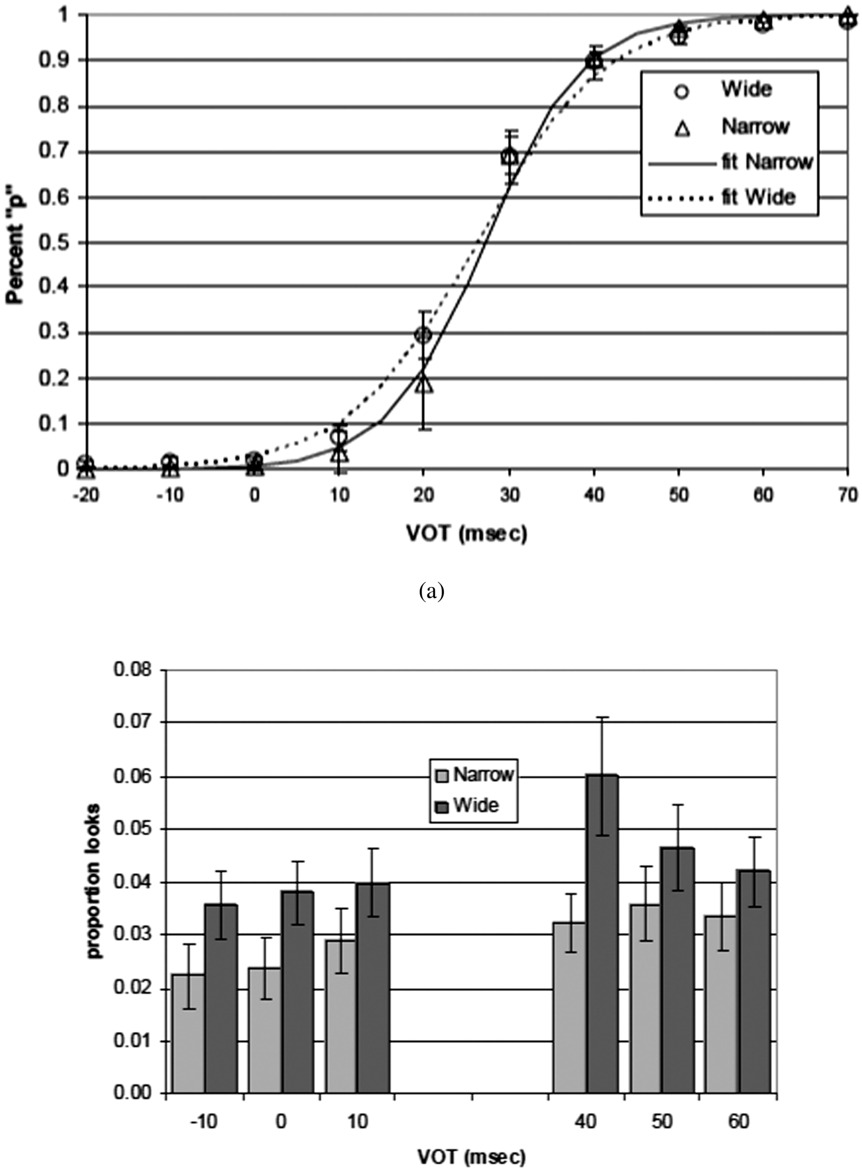
Adults were exposed to either the narrow or wide distribution in the context of a word-identification task using the Visual Word paradigm. In this paradigm, participants view a four-alternative picture selection display and hear instructions of the form "Click on the [word]." Their eye gaze is monitored as they perform this simple mouse-click task because these shifts in gaze have been shown to be reflective of the online decision process that eventuates in the final selection (i.e., the click of the mouse on one of the four pictures). Over the course of an hour of exposure to words that began with either the consonant /b/ or /p/ (as well as unrelated filler items), participants were exposed to exemplars of /b/-onset and /p/-onset words whose VOT was drawn from either the narrow or wide distribution.

As shown in Figure 3a, the mouse-click judgments supported the prediction that the narrow variance would induce sharper category boundaries—the slope of the identification function was steeper for participants in the narrow than in the wide condition. Moreover, as shown in Figure 3b, the probability that participants would move their gaze to the *incorrect* category (e.g., a look to peach when the exemplar was beach) was greater in the wide than in the narrow condition. This indicates that there was greater uncertainty about which word was presented in the wide condition, especially as the VOT value of that word approached the category boundary. It is important to note that the level of uncertainty reflected in these gaze data was estimated only from trials in which the participants accurately judged the identity of the word (i.e., their mouse-click was correct). Recall that the *means* of the two VOT categories were not altered—only the *variances* were manipulated. Overall, then, these data suggest that adults are rapidly taking into account the detailed distributional properties of the acoustic/phonetic input to which they are exposed and adjusting their category judgments accordingly. This mechanism of distributional learning is undoubtedly implicit in that participants are not aware of the differences in variance—and yet they must be capable of detecting the small differences in VOT that define the variances around the means.

The Clayards et al. (2008) study provides evidence for exquisite sensitivity to the distributional properties of speech in a laboratory context. But in the real world, listeners are confronted with much more variable input, in part due to hearing multiple talkers and even multiple dialects. How do listeners balance the need to correctly interpret subtle acoustic/phonetic differences *within* a given dialect and yet be flexible enough to adjust their interpretations across the much larger differences that are present *between* dialects? For example, when a speaker of American English says *boat* [bot] and a speaker of British English says *boat* [bowt], they intend to convey the same lexical item but instantiate it with different canonical pronunciations.

Jessica Maye addressed this question by determining whether listeners rapidly adjust their interpretation of dialect differences (Maye, Aslin, & Tanenhaus, 2008). Rather than explicitly training adults on specific lexical items (e.g., seeing a picture of a boat and hearing a talker say the word *boat* with a particular pronunciation), we asked whether mere exposure to a novel dialect would induce listeners to implicitly update their “pronunciation dictionary” for that dialect. We also asked whether this updating only applied to words that were contained in the input or whether it would generalize to unattested words that shared features of the novel dialect.

The basic design of the Maye, Aslin, et al. (2008) study was to have adults listen to a story—two chapters from *The Wizard of Oz*—spoken by a speech synthesizer. In Phase 1, the synthesizer used the standard pronunciation of American English. Thus, sentences were of the form “The wicked witch of



(a)

(b)

FIGURE 3 (a) Resultant labeling functions based on mouse-click responses. (b) Proportions of gaze directed to the nonselected (competitor) category choice.

the West." After this 45-min exposure to the story, participants performed a lexical-decision task with words in the story (e.g., *witch*), novel words (e.g., *brick*), and nonwords (e.g., **wetch*). Then in a second session at least a day later, the same participants listened to the same story but now pronounced in a different dialect. All of the front vowels were shifted such that /i/ became /I/, /I/ became /e/, and so on. It is important to note that the back vowels were *not* shifted (e.g., /a/, /o/, /u/). This feature of the design allowed us to determine if a shift of a portion of the vowel space was sufficient to induce a wholesale shift of the listener's entire vowel space. After this second exposure phase to the novel dialect, participants performed the same lexical-decision task. Note, however, that a subset of the formerly unacceptable nonwords (e.g., **wetch*) was now an acceptable pronunciation of words in the novel dialect (e.g., *wetch* was now the correct pronunciation of the word *witch* in the standard dialect).

The results from the lexical-decision test after the two exposure phases (standard and accented) indicated that participants were significantly more likely to judge a nonword in the standard dialect as a word after they had been exposed to the novel dialect, provided of course that the pronunciation conformed to a word in the novel dialect. Moreover, that shift to judging nonwords as words after exposure to the novel dialect generalized to test items that were not presented in the exposure phase. However, only front vowels underwent this adaptation effect; that is, participants did not extend their generalization to the unadapted back vowels. Finally, participants continued to judge test items as words even after being exposed to the novel dialect despite the fact that these test items would not have been pronounced in that form in the novel dialect (e.g., **witch* is not a word in the novel dialect because it would be pronounced as *weech*). Thus, participants are biased to *add* new pronunciations to their mental dictionary but not to *delete* old pronunciations, at least over the course of 45 min of exposure.

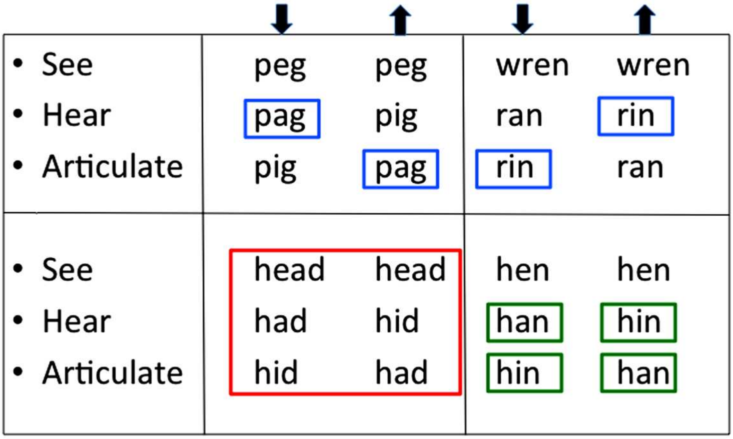
A natural question that arises from the Maye, Aslin, et al. (2008) study is whether these same processes of accent or dialect adaptation apply early in development when children are first learning words. Katherine White investigated this question in the context of known words and novel words (White & Aslin, 2011). We knew from prior work (Swingley & Aslin, 2000, 2002; White & Morgan, 2008) that infants could detect that a known word was mispronounced (e.g., that *gall* was intended to be *ball*). But what happens when infants hear a novel word? Do they treat it as a mispronounced version of a known word—presuming that the mispronunciation is similar to a known word—or as a new word to be added to their mental dictionary?

To address this question we presented 18-month-old infants with two pictures and measured to which picture they directed their gaze as a word was spoken. One picture depicted a known word and the other picture an unknown word. Replicating earlier studies, we found that when a mispronounced version of the known word was spoken, infants looked longer at the unknown object, suggesting that they detected the mispronunciation and by using the principle of mutual exclusivity (i.e., that each object has a single name) assigned the mispronounced word to the novel object. However, in a second group of infants the test phase was preceded by a labeling phase in which the known word was mispronounced in the context of a picture of the known object. This is analogous to having nonnative speakers of English demonstrate in an unambiguous context how they pronounce English words using their nonnative accent. For this group of infants, when confronted with the same test phase in which a picture of a known object and a novel object were present as the name of the known object was mispronounced, infants now looked longer at the known object. Moreover, when tested with known words that shared the same type of mispronunciation as was exhibited in the pretest labeling phase, infants generalized their acceptance of the mispronunciation to these words. These results are remarkably similar to those of Maye, Aslin, et al. (2008), although in a much simpler context (and without the separation of front vowels from back vowels).

In summary, the present review of three recent studies by my students illustrates that adults and infants are sensitive to the distributional properties of the ambient linguistic environment and utilize this information in a principled way to reach near optimal decisions about which phonetic category, and in turn which word, is being spoken in a given context. This process of distributional learning is also quite rapid, even in adults, suggesting that the phonological and lexical systems are undergoing continual updating based on both long-term and short-term weighting of distributional information.

SENSORY-MOTOR LEARNING IN SPEECH PRODUCTION

As noted earlier, there is a tight coupling between the acoustic/phonetic information to which listeners are exposed and the vocal-articulatory gestures that produce that information. Under ordinary circumstances, this renders any conclusions about the mechanism that links perception and production virtually impenetrable to experimental investigation. One wedge into this tightly coupled system is to perturb the auditory feedback that results from a given articulatory gesture. Houde and Jordan (1998) provided an early example of this perturbation paradigm by altering the formant structure of a single vowel that was repeatedly spoken and feeding that vocal output back to the listener in real time (less than 100 ms delay). What Houde and Jordan found was that adults rapidly adjust their articulatory gestures to compensate for the anomalous auditory feedback induced by the perturbation. This is analogous to what happens under conditions of prism adaptation in the visual-motor domain—participants adjust their motor command so that the intended target (spatial location in vision or phonetic category in



	⬇	⬆	⬇	⬆
• See	peg	peg	wren	wren
• Hear	pag	pig	ran	rin
• Articulate	pig	pag	rin	ran
• See	head	head	hen	hen
• Hear	had	hid	han	hin
• Articulate	hid	had	hin	han

FIGURE 4 Design of the Frank et al. (2014) perturbation study. Arrows indicate direction of vowel shift (color figure available online).

speech) is achieved and show a negative aftereffect when the perturbation is removed.

In Austin Frank's dissertation, we were interested in whether the lexicon plays a role in how this auditory-motor adaptation process operates (Frank, Aslin, Jaeger, & Tanenhaus, 2014). In Houde and Jordan (1998), every participant repeated a single vowel, /I/, with half undergoing a perturbation toward /i/ and the other half toward /e/. A similar design was used in Frank et al., but now every participant was reading a list of words that contained a *variety* of vowels. On each trial, participants saw the text of a word in English and attempted to produce it. Prior to any perturbation, what they saw would lead them to produce the target word and they would hear themselves saying that word (i.e., all three levels of representation—text, motor command, and auditory feedback—would be isomorphic). However, when the perturbation was implemented, these three levels became uncoupled. As shown in Figure 4, the list of words was carefully chosen so that they formed triples in which all, some, or none of the members of the triples were words in English. Consider the bottom row in this figure. In the left panel, all three members of the triple are words in English (regardless of an upward or downward shift in the first formant of the vowel), whereas in the right panel none of the members of the triple are words. Crucially, in the top row, you can see that, depending on the direction of the vowel shift, we have uncoupled whether the auditory target or the motor command that must adapt to the perturbation is a word or a nonword in English.

The results of this study show a strong role for the lexicon in the *magnitude* of adaptation to the perturbation. The greatest adaptation occurred when *neither* the auditory target nor the motor command conformed to a word in English (i.e., the lower right panel in Figure 4). No significant adaptation occurred in any condition where there was articulatory competition (i.e., when the new motor command would have produced a word in English). And when there was no articulatory competition but there was auditory competition (i.e., the new auditory target was a word in English), the degree of adaptation was intermediate. These results not only show that the lexicon influences the magnitude of adaptation to speech perturbation but also that this recalibration of the auditory-motor system involved in speech production is remarkably flexible—on a trial-by-trial basis participants did not know which word would appear on the screen, yet they adjusted their articulations within the 100-ms period as the vocal gesture was being implemented.

SUMMARY AND CONCLUSIONS

The overall theme of the four studies briefly reviewed in this article is that listeners and speakers maintain a close coupling between their perception of phonetic categories and the articulatory gestures required to produce these categories. This coupling is under adaptive control so that both slow changes in the overall distributional properties of the ambient linguistic input (e.g., dialect shifts) and rapid changes associated with switching among different talkers are seamlessly compensated for to ensure effective communication. Once a mature lexicon has been acquired, both the acoustic and articulatory similarities among words influence the ease with which the coupling between speech perception and speech production can be accomplished. These issues of sensory and motor coupling and adaptive plasticity were an important part of my training with Herb Pick in the mid-1970s, and it is to him and his intellectual rigor and unfailing enthusiasm for science that I dedicate this article.

ACKNOWLEDGMENT

This article is an edited version of a talk given at a Festschrift in honor of Herbert L. Pick, Jr., on June 5, 2012, in Minneapolis, Minnesota.

FUNDING

Preparation of this article was made possible, in part, by support from National Institutes of Health Grants HD-037082 and DC-05071.

REFERENCES

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

Frank, A. F., Aslin, R. N., Jaeger, T. F., & Tanenhaus, M. K. (2014). *On-line integration of linguistic, motor, and perceptual information in language production.* Manuscript in preparation.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279,* 1213–1216.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5,* 831–843.

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

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.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11,* 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95,* B15–B26.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33–B42.

Pisoni, D. B., & Tash, J. (1974) Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15,* 285–290.

Raphael, L. J., Borden, G. J., & Harris, K. S. (2007). *Speech science primer: Physiology, acoustics, and perception of speech.* Baltimore: MD: Lippincott, Williams & Wilkins.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 75,* 1–20.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and word-form representations of 14-month-olds. *Psychological Science, 13,* 480–484.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7,* 49–63.

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21,* 221–226.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents in toddlers. *Developmental Science, 14,* 372–384.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language, 59,* 114–132.