Cognitive, Affective, & Behavioral Neuroscience
2007, 7 (1), 53-73

Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps

Gautam K. Vallabha and James L. McClelland
Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania
and Carnegie Mellon University, Pittsburgh, Pennsylvania

The influence of a native language on learning new speech sounds in adulthood is addressed using a network model in which speech categories are attractors implemented through interactive activation and Hebbian learning. The network has a representation layer that receives topographic projections from an input layer and has reciprocal excitatory connections with deeper layers. When applied to an experiment in which Japanese adults were trained to distinguish the English /r/–/l/ contrast (McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002), the model can account for many aspects of the experimental results, such as the time course and outcome of the learning, how it varies as a function of feedback, the relative efficacy of adaptive and initially easy training stimuli versus nonadaptive and difficult stimuli, and the development of a discrimination peak at the acquired category boundary. The model is also able to capture some aspects of the individual differences in learning.

One of the key issues that a theory of perception must address is the effect of prior experience. Newborn infants initially have the ability to distinguish a rich variety of linguistic contrasts, but by about six months of age, their ability has begun to attune to their native language (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). This reshaping of the perceptual space presumably allows the developing infants to recognize their language sounds more effectively and may contribute to the warping of perceptual representations toward category prototypes (the *perceptual magnet effect*; Kuhl, 1991, 2000) and to the sharpening of sensitivity at category boundaries (the *categorical perception effect*; Liberman, Harris, Hoffman, & Griffith, 1957). However, there is the suggestion that such reshaping may have the side effect of hindering the acquisition of language distinctions later in life (Flege, 1995). In this paper, we explore a mechanistic account of how first-language speech acquisition can influence the later acquisition of second-language speech sounds.

Our exploration is guided by the assumption that speech learning is an instance of more general architectural and computational principles. We consider four such principles and evaluate their usefulness in addressing data from an experiment on the acquisition of a nonnative speech contrast in adulthood. Specifically, we use the principles to develop a computational model of the pattern of successes and failures in learning the American English /r/ and /l/ by adult native speakers of Japanese (McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). Models of perceptual learning often attempt to reproduce only the end state of learning, but frequently, this does not place sufficient constraints on the models (Damper & Harnad, 2000; Edelman & Intrator, 2002). In order to impose additional constraints, we evaluated our model against several measures over the time course of learning and under different training conditions and explored the extent to which the principles can account for a wide range of findings.

The article is organized as follows. First, we give an overview of the R/L problem and of the McCandliss et al. (2002) results, pointing out aspects of the results that appear puzzling in the absence of an explicit mechanistic explanation. Next, we introduce and justify the principles underlying our model and present an abstract and simplified version to illustrate its basic properties. We then set out our model of the McCandliss et al. data and present the modeling results together with the experimental results. Finally, we evaluate the successes and limitations of the model and the consequent implications for the architectural principles.

The R/L Problem

The Japanese language does not have distinct /r/ and /l/ phonemes, and native speakers of Japanese have extreme difficulty distinguishing between the /r/ and /l/ sounds in American English (AE), particularly when they occur in syllable-initial positions (Logan, Lively, & Pisoni, 1991). When presented with an /r/–/l/ continuum in which the $F3$ onset changes systematically from 1362 Hz (for /r/) to 3698 Hz (for /l/), AE listeners show a fairly sharp discrimination peak at the boundary between the two categories, whereas Japanese listeners show no such boundary even after many years of English experience (Miyawaki et al., 1975). In place of the English /r/ and /l/ sounds, Japanese

G. K. Vallabha, vallabha@cnbc.cmu.edu

has an apico-alveolar tap /ɾ/ whose categorization appears to depend primarily on the second formant (Iverson et al., 2003; Lotto, Sato, & Diehl, 2004). Consequently, when asked to label AE /r/ and /l/ sounds by a native speech category, Japanese listeners usually map both classes of sounds to the Japanese /ɾ/, sometimes prefixing the unrounded high back vowel /ɯ/ to it (Guion, Flege, Akahane-Yamada, & Pruitt, 2000).

There have been several approaches to training Japanese listeners to perceive the /r/–/l/ distinction (see, e.g., Lively, Logan, & Pisoni, 1993; Strange & Dittmann, 1984). The patterns of acquisition are varied and complex, but there are also some common characteristics (see Iverson, Hazan, & Bannister, 2005, for a comparative study). Specifically, learning is rarely complete and learners usually do not match the performance of the native speakers; there are considerable individual differences in learning (Takagi, 2002); learning is often facilitated by the use of easy–hard procedures in which contrasts are initially easy but progressively become difficult (see, e.g., Jamieson & Moroson, 1989); generalization is facilitated if there has been some stimulus variability (Lively et al., 1993); and finally, learners show a general bias toward L-labeling after training (see Figure 6 of Iverson et al., 2005). For concreteness, in the present article, we focus on the training conducted by McCandliss et al. (2002). In addition to showing several of the characteristics enumerated above, the study explores two phenomena not often addressed in /r/–/l/ training studies: the effectiveness of training without feedback, and the effect of category learning on discrimination. Consideration of these phenomena allows the issue of /r/–/l/ learning to be connected to phenomena in other domains. For example, the question of learning with and without feedback has been extensively explored in visual learning (Herzog & Fahle, 1998; Petrov, Dosher, & Liu, 2005; Rosenthal, Fusi, & Hochstein, 2001; Tarr & Cheng, 2003). Furthermore, visual learning often results in changes in discriminability (e.g., Goldstone, 1994) that are similar to those observed in speech perception (e.g., Kuhl, 1991).

The McCandliss et al. (2002) study originated from the proposal of McClelland, Thomas, McCandliss, and Fiez (1999) that speech perceptual learning may depend on an unsupervised Hebbian learning process. In this proposal, the presentation of a speech sound elicits a perceptual representation that is viewed as a pattern of activation over a population of neuron-like processing units. Hebbian learning (strengthening of the connections between units contributing to this pattern of activation) tends to strengthen the tendency for the speech sound to elicit the same perceptual representation in the future. For the Japanese listeners, /r/ or /l/ both elicit an /ɾ/-like representation, and the resulting Hebbian learning causes further entrenchment of this tendency. This perspective makes two predictions. One is that the learning may be facilitated by using exaggerated tokens of /r/ and /l/ to elicit distinct representations. Once the representations gain a foothold, then Hebbian learning should strengthen the contrast between them and eventually result in robust perceptual learning. The second prediction is that since the establishment and strengthening of representations is entirely unsupervised, learning should not depend on outcome information. McCandliss et al. (2002) tested these predictions by training Japanese listeners in different conditions. Since our model attempts to account for these results, we shall describe the experiment and results in some detail.

Summary of Experimental Results From McCandliss et al. (2002)

McCandliss et al. (2002) first recorded /r/ and /l/ words spoken by a native speaker of English and acoustically manipulated them to generate continua from exaggerated /l/ to exaggerated /r/. These stimuli were then used to train Japanese subjects in four conditions—fixed versus adaptive training stimuli, crossed with the presence versus absence of feedback. In the *fixed* conditions, subjects always heard the same two intermediate (and initially confusable) tokens of /r/ and /l/ and had to label them as "r" or "l." In the *adaptive* conditions, subjects initially heard exaggerated tokens that were easily identifiable; when the participant identified eight successive tokens correctly, the /r/ and /l/ tokens moved closer to each other. In the *feedback* conditions, the subjects received immediate visual feedback on the correctness of each response; in the *no-feedback* conditions, this visual feedback was omitted. There were 8 subjects in each condition, 4 trained with a *rock–lock* continuum and 4 with a *road–load* continuum. The training was conducted over 3 days (with daily sessions of 500 trials), with half of the subjects given an additional three sessions of training.

The learning of the subjects was evaluated in four ways: (1) performance on probe stimuli that were periodically presented to the subject during the training, (2) categorization on trained and untrained R/L continua, (3) same–different discrimination of pairs with fixed interstimulus distance ("slide test"), and (4) same–different discrimination of pairs with increasing interstimulus distance ("expand test").

The following are the main experimental results that we seek to explain. (1) Adaptive training without feedback was effective in inducing perceptual learning (Figure 6D). (2) Feedback markedly improved the learning and also inverted the relative efficacy of the conditions (without feedback, adaptive training was better; with feedback, fixed training was better). The training efficacy (as measured by the rate of learning) may be rank ordered as follows: fixed-with-feedback, adaptive-with-feedback, adaptive-without-feedback, fixed-without-feedback (Figure 7B, Figure 6D). (3) The learning was not stimulus specific. After training, subjects (particularly in the feedback conditions) showed improved categorization on the untrained R/L continuum (Figure 6D). (4) The improvement in R/L classification was paralleled by an increase in discriminability for stimulus pairs that straddled the category boundary (shown by the slide test, Figure 9D). Discriminability also improved for stimulus pairs that straddled the category boundary with greater interstimulus distances (shown by the expand test, Figure 10B). (5) There were marked individual differences in the no-feedback training. Some subjects learned quickly even though they were in the fixed condition, whereas others learned slowly even though they were in the adaptive condition. (6) In all conditions, there was a strong bias toward labeling the stimuli as "L"; that is, the posttraining category boundary was closer toward the "R" end of the L–R continuum than in native English speakers (Figure 6D). (7) Learning in the fixed-without-feedback condition was very slow, hardly improving between the pre- and posttraining categorization curves (Figure 6D).

The results of the experiment are partially consistent with the simple Hebbian account, but also demonstrate that the account is incomplete. These results supported the Hebbian analysis that eliciting separate representations can induce category learning, even in the absence of feedback about response accuracy. However, there was a clear effect of feedback. In short, although learning can take place without feedback, the speech perceptual system can take advantage of feedback when it is available.

Note that the first six results are generally consistent with those observed in other perceptual learning and /r/–/l/ training studies (e.g., Iverson et al., 2005) and are therefore of general theoretical interest. The last result (about slow learning in the fixed-without-feedback condition) has not been explicitly noted elsewhere to our knowledge, but it may have real-world validity, given that the majority of Japanese speakers learning English are exposed to normal (i.e., unexaggerated) /r/ and /l/ sounds without consistent feedback. Thus, the lack of learning in the fixed-without-feedback condition may shed light on the larger causes of difficulty in /r/–/l/ learning. Consequently, our goal in the present work was to develop a model of perceptual learning for speech that can account for perceptual learning without feedback, the role of feedback when it is available, and the conditions that lead to the lack of learning. The formulation of the model will be constrained by two factors: The *empirical* constraint is the detailed pattern of findings in the McCandliss et al. (2002) experiment (described above), and the *theoretical* constraint is a general conception of language learning and neural architecture. In the following section, we describe the theoretical constraint in more detail.

THEORETICAL APPROACH

General Conception

We first give our overall conception of the initial speech learning process and how it affects second language acquisition and then describe the general principles governing the formulation of the model itself.

We work within the context of the idea that an infant's perceptual space is initially quite plastic and subject to structuring through experience (Kuhl, 2000). From birth, infants are exposed—without explicit labels—to sounds of the ambient native language. They may receive correlated inputs (for example, the sight of the facial movements of the speaker), but for simplicity, we will not include such influences. As the experience accumulates, perceptual representations that are activated most frequently (such as those corresponding to prototypical sounds of the native language) come to function as "perceptual magnets" (Kuhl, 1991) or "attractors" (Flege, 1995). That is, there is a graded category structure with sounds near the centroids of the clusters that produce stronger perceptual representations than those at the edges of these clusters (this is similar to the mechanism used by Rosenthal et al., 2001, to address unsupervised visual category learning; see also Anderson, Silverstein, Ritz, & Jones, 1977). Inputs near an attractor point tend to be distorted in the direction of the attractor, making stimuli near the same attractor more similar to each other. These perceptual attractors are strengthened throughout childhood and gradually become strongly entrenched. If an adult listener hears new sounds (e.g., from a second language) that are near the center of an existing attractor, the distortion produced by the attractor will reduce discriminability between the sounds and retard acquisition of the new distinction. Consequently, age-dependent effects on perceptual learning are taken to be gradual and experience dependent rather than strictly dependent on biological factors, such as puberty (Flege, 1992; White, 2001).

It is into the context established by such experience that an adult native speaker of one language would come to a perceptual learning experiment such as that of McCandliss et al. (2002). In the present article, we take the view that mechanisms similar to those operating to establish perceptual attractors through development are also operating to influence the course of learning in the experiment. Within the context of the above conception, we go on to explore the question of how feedback may influence the perceptual learning process, by either modulating or supplementing the mechanisms that allow essentially Hebbian unsupervised learning processes to shape the attractors affecting perceptual representations.

General Principles Governing Model Formulation

Our model is intended to instantiate this general conception in a simplified framework in which the actual dynamics of speech sounds are ignored and sounds are treated as points in a simplified input space spanning two dimensions. This simplification allows us to explore whether the problem of speech perceptual learning may be formulated in terms of the following general principles that are widely applicable in other domains of cognition.

Processing engages attractors in a multilayer network via interactive activation and competition. The perceptual system is assumed to be organized around a set of processing principles articulated in McClelland (1992) and embodied in several models of perception and perceptual identification (e.g., the stochastic interactive activation model of McClelland, 1991, and Movellan & McClelland, 2001, and the ART framework of Carpenter & Grossberg, 1987, and Goldstone, Steyvers, & Larimer, 1996). The principles specify that processing takes place in a network consisting of several interacting layers with inhibitory connections within a layer and excitatory connections between layers. The activation of each unit is a graded nonlinear function of the net input, and the activation propagates continuously in time across the layers. The network is intrinsically stochastic; that is, random noise perturbs the inputs to units at each time step.

Perceptual learning increases the strength of the attractors established by the interactive dynamics. This increase in strength may be instantiated as an increase in the neural activity to stimuli similar to those experienced frequently. Support for this assumption comes from a series of studies by Recanzone, Merzenich, and Schreiner (1992a, 1992b). In their experiment, monkeys were trained to discriminate between the frequencies of vibrotactile stimuli applied to their fingertips. After the training, the stimulation of the trained skin resulted in temporally synchronous neural activity over a larger cortical area.[1] This result is consistent with some imaging studies of humans (e.g., Pleger et al., 2003). Of specific relevance to our own work, Callan et al. (2003) and Callan, Jones, Callan, and Akahane-Yamada (2004) explored this issue in Japanese subjects who were trained to discriminate /r/ and /l/ using the materials of Logan et al. (1991). After training, their subjects exhibited increased activity in auditory and speech motor areas while performing an /l/–/r/ classification task.

Learning occurs in a Hebbian manner both in the presence and absence of correlated information, indicating the category membership of the input. This assumption is consistent with behavioral evidence (McCandliss et al., 2002; Rosenthal et al., 2001) and other modeling approaches (Carpenter & Grossberg; 1987; Petrov et al., 2005; Rumelhart & Zipser, 1985). Furthermore, Hebbian processes have been demonstrated to be sufficient for synaptic change in the auditory cortex. For example, Ahissar, Abeles, Ahissar, Haidarliu, and Vaadia (1998) monitored a neuron A in the auditory cortex of a monkey while it was being trained on a discrimination task and electrically stimulated neuron B whenever A was active. After the training, there was increased correlation between the activities of the two neurons, suggesting the presence of Hebb-like synaptic change (see Syka, 2002, for a review of related results).

Perceptual processes and representations occur within topographic maps. There is substantial evidence both for the existence of topographic maps and for their plasticity (Buonomano & Merzenich, 1998; Kohonen, 1993). Such representations allow efficient representations, with spatially bounded receptive fields allowing for distributed sparse coding and the overlap between receptive fields allowing for noise tolerance and generalization (Idiart, Berk, & Abbot, 1995; Poggio, 1990). Moreover, objects in the world are cohesive along dimensions such as time, space, and frequency. A topographic map "projects" this cohesiveness directly into the perceptual system and can thereby guide its development, as demonstrated by models of the maturation of the early visual pathways (Linsker, 1986; Sirosh & Miikkulainen, 1997) and by models of somatotopic map reorganization (Grajski & Merzenich, 1990; Sutton, Reggia, Armentrout, & D'Autrechy, 1994).

In our model, the above principles are instantiated in a multilayer network with bidirectional inhibitory connections within a layer, bidirectional excitatory connections between layers, and Hebbian update of the between-layer connections. The bidirectional connections implement attractors, a mechanism common to many earlier models, such as interactive activation models of written and spoken word perception (McClelland & Elman, 1986; McClelland & Rumelhart, 1981) and models of associative memory (e.g., Anderson et al., 1977). Our model also shares many features with Grossberg's ART networks (Carpenter & Grossberg, 1987; Grossberg, 1988; see also Hoshino, 2002), with assumptions based on Grossberg's for learning in the absence of supervision. There are differences, however: ART models do not often employ topographic maps, and we do not employ some features of ART models, such as the use of a mismatch detection mechanism.

We present our model in two stages, each instantiated in a separate simulation. Simulation 1 uses an abstract and simplified version of the model to illustrate many of its basic properties as well as how certain characteristics of perceptual learning would arise within models of this type. Simulation 2 uses an elaborated version of the model to address the acquisition of the /r/–/l/ contrast by Japanese adults, as was investigated in the McCandliss et al. (2002) experiment.

SIMULATION 1
Aspects of the Nature and Acquisition of Categories of Speech Sounds

Simulation 1 illustrates how our model addresses—at an abstract level—several putative aspects of the nature and acquisition of categories of speech sounds. These aspects all play key roles in our account (in Simulation 2) of the acquisition of the /r/–/l/ distinction in Japanese adults. Although there has been disagreement about some of these aspects, they have all been argued for by others, and we list some of the main protagonists of each point. (1) Perceptual categories can be acquired in an unsupervised manner (Kuhl et al., 1992; Rosenthal et al., 2001). (2) Membership in the resulting perceptual categories can be graded; that is, some stimuli may be better exemplars of the category than others (Kuhl et al., 1992; Miller, 1994; Oden & Massaro, 1978). (3) There appears to be an inverse relation between the goodness of category exemplars and their discriminability (the *perceptual magnet effect*; Kuhl, 1991). (4) Discriminability is usually maximal at category boundaries (cf. the *categorical perception effect*; Liberman et al., 1957).

Description of Model

Architecture. The network consists of three one-dimensional layers, designated L1, L2, and L3 (Figure 1A). L1 is topographically connected to L2; each L2 unit has the strongest connection to the L1 unit directly below it and exponentially weaker connections to L1 units that are further away. Specifically, the weight between L2 unit *a* and L1 unit *b* is given by $\exp(-(a - b)^2 / \beta_{21})$, with $\beta_{21} = 15$. The incoming L1 weight vector to each L2 unit is normalized to have unit magnitude, and the input space does not wrap around. L2 is bidirectionally connected to L3, with the L2→L3 weights initialized from a uniform distribution between 0 and 0.03 [notated as *Uniform*(0, 0.03)] and the L3→L2 weights from *Uniform*(0, 0.0005). The asymmetry is needed to prevent the recurrent excitation from exerting an influence too early in learning, which could cause the network to lock on to spurious attractors. Finally, each L2 unit is connected to itself with a weight of $+1.0$ and to all others

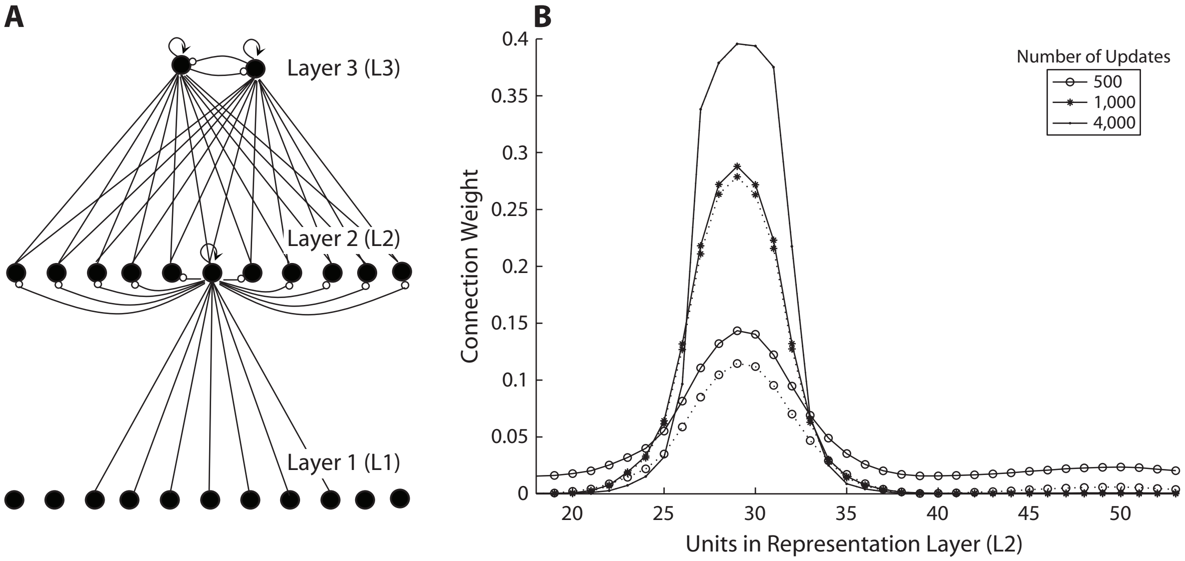

Figure 1. (A) The network architecture for Simulation 1 (with 80 units in L1 and L2 and 2 units in L3) depicting topographic connectivity from L1→L2, bidirectional full connectivity between L2 and L3, and self-excitation and lateral inhibition by units in L2 and L3. (B) The weights from L2 to L3 unit 1 after 500, 1,000, and 4,000 updates (solid lines) and the corresponding recurrent weights (dotted lines).

in its layer with a weight of −0.2; L3 units have similar connectivity except that the inhibitory weights are −2.0.

Network dynamics. Each unit in *k* in L2 and L3 gradually accumulates its incoming activity (its "net input" net_k) and converts it to an instantaneous firing rate (its "output activity" a_k). The net input tends to decay to zero and is also subject to noise during the integration. The output activity increases monotonically with nonnegative net input and asymptotes toward a maximum value of 1.0. These values were regulated by the following differential equation:

$$dnet_k = \left(-net_k + \sum_j wscale_{kj} \cdot w_{kj} a_j\right) \cdot dt + \xi_{\text{net}} \sqrt{dt},$$

$$a_k = \max\left[0, \tanh\left(net_k \cdot gain_{\text{act}}\right)\right],$$

where w_{kj} are the incoming weights, ξ_{net} is the integration noise [$\xi_{\text{net}} \sim N(0, \sigma_{\text{net}})$], and $gain_{\text{act}}$ is the gain of the activation function. $wscale_{kj}$ is a "weight scaling" parameter, set to 1 for L1→L2 and L2→L3 weights and to 5 for L3→L2 weights. We assume that each L3 unit is a member of a group of neurons with similar incoming and outgoing projections, as would be the case if L3 is a topographic map. The recurrent excitation to an L2 unit *k* is the summed input from all neurons in that group, which we approximate by scaling up the influence of the single active L3 unit. Before each input was presented, Gaussian noise was added to the input pattern, $I_i^{\text{noisy}} = I_i + \xi_{\text{input}}$, where $\xi_{\text{input}} \sim N(0, \sigma_{\text{input}})$. The activations of the input units were clamped to the noisy input pattern for 30 time steps, and the network was allowed to settle with $dt = 0.2$, $\sigma_{\text{input}} = 0.02$, $\sigma_{\text{net}} = 0.2$, and $gain_{\text{act}} = 0.5$.

Weight update. After the settling was completed, the weights between L2 and L3 were updated using a Hebbian rule: $\Delta w_{jk} = \eta\, a_j\, a_k$, with $\eta = 0.02$. For simplicity, we turned off the learning between L1 and L2. However, simulations in which L1–L2 learning was allowed produced similar results. Following the update, the weights were multiplicatively normalized in a graded manner; that is, for small weight-vector magnitudes there was no normalization, but as the weight vector magnitude increased, the normalization was applied with greater force (if the weight vector magnitudes are always fully normalized, then the initial random configurations would get reinforced and prevent further learning). Let *w* be the weight vector before the weight update and Δ*w* the change in the weight vector. Then m_{new}, the magnitude of the new weight vector, is defined to be:

$$m_{\text{new}} = [1 - f(|w|)] \cdot |w| + f(|w|) \cdot |w + \Delta w|,$$

where $f(x)$ is a function that decreases from 1 to 0 over the range $[0, M]$, and M is the maximum magnitude of the weight vector. We chose $M = 1$ and $f(m) = \max[(M - m)^3, 0]$, but the learning is not sensitive to the value of the exponent and only requires that it be ≥ 1. For L2→L3 projections, the weight vectors to be normalized are the sets of *incoming* weights. For L3→L2 projections, the weight vectors are the sets of *outgoing* weights. This configuration forces the L2→L3 and L3→L2 weights to be approximately symmetric and is required to achieve the attractor behavior (Grossberg, 1988).

Inputs. Each input stimulus was a bump of activity on the input layer. The center of the bump was the "input location" and could vary continuously between 1 and 80. For an input location *x*, the activity of input unit *i* was a Gaussian bump centered at *x*, given by $\exp(-(i - x)^2 / \beta_{\text{input}})$, with $\beta_{\text{input}} = 2$. The resulting input pattern was normalized to have a magnitude of 1.0. The input locations were drawn from two Gaussian distributions, $N(29, 3)$ and $N(51, 3)$, with 400 locations from each distribution. The input pat-

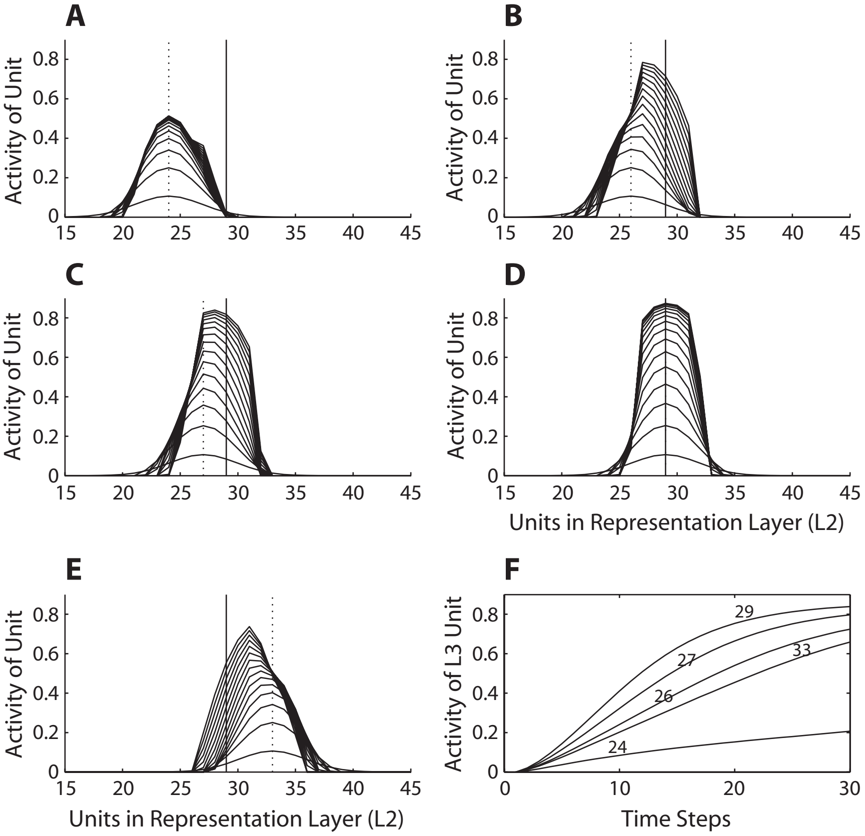

Figure 2. The attractor dynamics of the network in Simulation 1 after 4,000 updates (the input and integration noise were turned off to provide a clearer view of the network's tendencies). (A) The development of L2 activity for an input centered at location 24 (vertical dotted line). The center of the category is at location 29 (vertical solid line). The rising curves indicate L2 activity at successive points in time. (B–E). The L2 activity for inputs at locations 26, 27, 29, and 33. (F) The activity of the winning L3 unit for the same set of inputs. The inset number is the location of the input.

terns were presented to the network in random order 4,000 times, and the between-layer weights were updated after each presentation.

Results and Discussion

Each time an input is presented to the network, it results in a pattern of activity over L2. The L3 unit that is most strongly activated by this pattern suppresses the other L3 units, and the resulting weight update increases its sensitivity to that particular pattern. Through this process of competitive learning (Rumelhart & Zipser, 1985), the L3 units divide up the L2 space so that each L3 unit becomes most responsive to one input cluster; thus, the L2↔L3 weights eventually reflect the structure of the categories (Figure 1B). Consequently, whenever an L3 unit is activated, it recurrently excites the prototypical L2 representation for that category, which in turn further excites the L3 unit, and so on; this reverberating activity between L2 and L3 instantiates a *perceptual attractor*.

Figure 2 illustrates the attractor dynamics after training. When the stimulus is outside the category (Figure 2A), the L2 activity is relatively unbiased. When the stimulus is closer to the category (Figure 2B), the recurrent excitation from L3 increases the total amount of L2 activation. In addition, the L2 activity pattern ("perceptual representation") becomes biased toward the center of the category. However, the original input still exerts an influence, so the representation does not completely converge toward the category center. The attractor set up by the recurrent excitation is not a fixed state that is the same for all inputs, but is more like a flexible schema (Rumelhart, Smolensky, McClelland, & Hinton, 1986) that is adapted toward the current input. The resonance between L2 and L3 is self-sustaining, and if the input was turned off during the settling process, the L2 activity would shift entirely toward the center of that input's category.

The attractor dynamics have two consequences. First, the "category goodness" of a stimulus (defined as the activity of the winning L3 unit) increases gradually toward the center of the category (Figure 3A). Second, the discriminability of adjacent stimuli (defined as the Euclidean distance between the corresponding L2 activity patterns) is inversely proportional to their category goodness (Figure 3B). The discrimination minima occur because stimuli near the center of a category engage the same attractor; the maxima occur because stimuli within a category evoke greater levels of

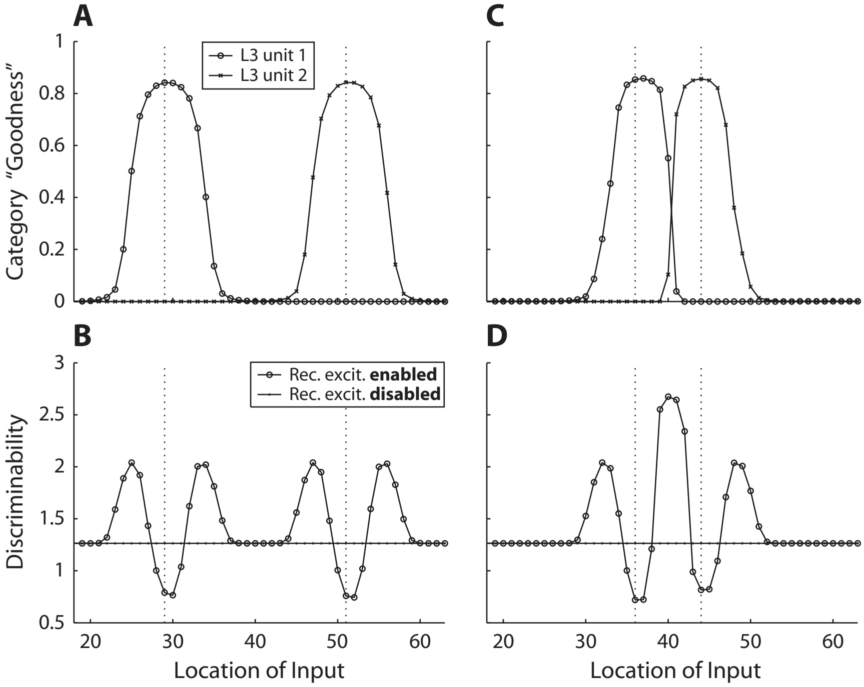

Figure 3. Average category goodness and discrimination curves in Simulation 1 after 4,000 updates. The curves are averaged over 10 runs. (A) The category goodness—that is, the activation of the L3 units—as the input is varied over the input space. The dotted vertical lines mark the centers of the input categories (29 and 51). (B) The discriminability between pairs of stimuli, with recurrent excitation either enabled or disabled during test. The discrimination value for input location x is the average Euclidean distance between the L2 patterns at time step 30, evoked by stimuli at locations $x-2$ and $x+2$. (C, D) The goodness and discrimination when the input categories are close to each other (centers at 36 and 44).

L2 activity than stimuli outside the category, and this difference in activity level increases discriminability between stimuli on different sides of a category boundary (note that these effects are entirely abolished if the recurrent excitation is disabled). Figures 3C and 3D show that the relation between category membership and discrimination persists even when the input categories are close together.

In summary, we have demonstrated that a network with recurrent excitation from a "category layer" to a "representation layer" can show phenomena such as unsupervised category acquisition, graded category membership, gradual assimilation toward an attractor, and discriminability that is inversely proportional to the category goodness. The time course of the network's activity is also consistent with behavioral data that speech percepts tend to become more categorical as the interstimulus interval is increased (Pisoni, 1973).

SIMULATION 2
Model of the McCandliss et al. (2002) Experiment

Simulation 1 indicates that an interactive Hebbian model can capture several putative aspects of the acquisition and processing of speech categories and also illustrates how the development of activity in the representation layer contributes to the network's performance. In the following simulation, we use the same architecture to account for the perceptual learning in McCandliss et al. (2002)—in particular, the pattern of learning with the presence versus absence of feedback and with fixed versus adaptive stimuli. We also use Simulation 2 to illustrate our conception of why second-language acquisition is difficult, how this difficulty relates to the McCandliss et al. training, and which factors underlie intersubject differences in the training.

Before proceeding further, a terminological caution is necessary. In the literature, the term *feedback* is used to refer to either outcome information provided to the subject or recurrent connectivity in a neural network. We use it only in the former sense and use the term *recurrent excitation* to refer to the latter.

Phases of Learning and Basic Model Architecture

The network to accomplish the McCandliss task is a two-dimensional version of that in Simulation 1 (Figure 4A), with L1 and L2 implementing a topographic map over two acoustic features that are most relevant for distinguishing /r/ and /l/ (Figure 4B) and L3 detecting clusters in the L2 activity distributions.

The simulation consists of two phases. In Phase 1 ("Japanese environment"), the network is exposed to unlabeled instances of the Japanese alveolar tap /ɾ/ and velar approximant /ɰ/. These two Japanese sounds are acoustically the most confusable with AE /r/ and /l/; thus, by training the network with these sounds, we attempted to capture the warping of the Japanese perceptual space near /r/ and /l/. We proceed from the assumption that exemplars of the Japanese /ɾ/ are highly variable (particularly along *F*3) and are spread out over a region that for an English listener would include both the /r/ and /l/ categories (Lotto, Sato, & Diehl, 2004). The Japanese-trained network eventually develops an attractor so that stimuli within the /ɾ/ category (and hence /r/ and /l/ stimuli) are less discriminable from each other. In Phase 2 ("R/L training"), the Japanese-trained network is trained to categorize AE /r/ and /l/ using the training conditions in McCandliss et al. (2002)—that is, fixed stimuli with and without feedback and adaptive stimuli with and without feedback.

Two caveats are in order regarding the aforementioned architecture. First, L2 and L3 should be thought of as a single perceptual system. In particular, we emphasize that we do not think of the L3 units as localist phoneme representations, but as approximations to a distributed (possibly topographic) representation. Second, while actual speech is time varying, the input stimulus to the model is static. Moreover, the space of *F*2 and *F*3 onsets is just one slice through a much higher dimensional space (comprising, for example, the other formants and the overall amplitude envelopes of the sounds). The localist L3 representation and the static, low-dimensional inputs are therefore simplifications adopted for the sake of allowing the exploration of more general principles.

Model Elaboration

The general principles, learning phases, and architecture that were previously discussed provided the starting place for our development of a model of the findings of the McCandliss et al. (2002) experiments. When we turned attention to applying the model to these findings, additional issues arose that required careful consideration. We will consider these issues in turn.

Differentiation of network architecture into fast- and slow-learning systems. The first issue concerns the difference in timescales between the initial native-language exposure and the duration of the McCandliss et al. (2002) experiment. The Japanese subjects had experienced Japanese language sounds intensively for many years, so it is unlikely that their perceptual systems were fluctuating on a day-to-day basis or that they could be fundamentally restructured over a brief period. Yet, their perceptual abilities were modified after only three days of training in the McCandliss et al. experiment. So how should the relation between the long-term exposure and the experimental learning be understood?

One approach to this problem of "catastrophic interference" is to assume that L3 has a large pool of unused units; when a stimulus is sufficiently distant from all existing categories, a new unit is recruited to represent that stimulus (see, e.g., Carpenter & Grossberg, 1987). However, this approach is difficult to apply when (as in the case of /r/ and /l/ and the Japanese tap) the new stimulus cat-

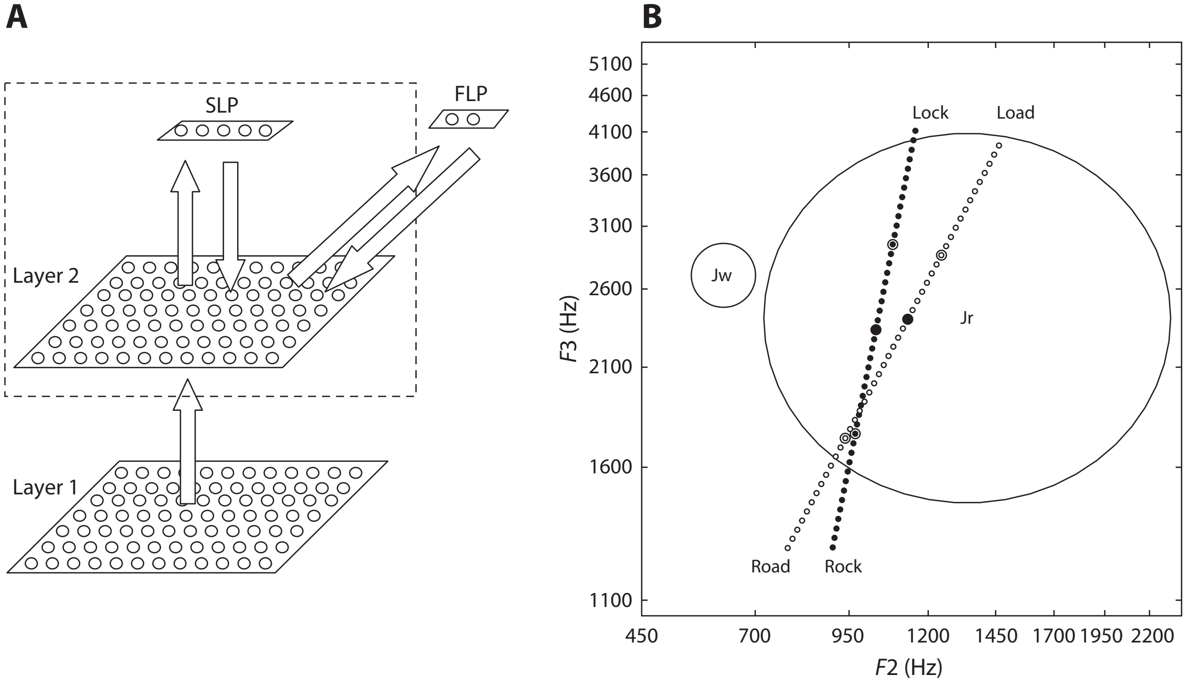

Figure 4. (A) The network architecture for Simulation 2. L1 and L2 are 20 × 20 grids of units. L2 is bidirectionally connected to the slow learning pool (SLP; with five units, used in Phase 1 training) and to the fast learning pool (FLP; with two units, used in Phase 2 R/L training). All between-layer connections are excitatory. Each unit in L2, SLP, and FLP excites itself and inhibits all the other units in its layer. (B) The input space of the network. The two large circles indicate the distribution of the two Japanese categories ("Jr," Japanese apico-alveolar tap; "Jw," Japanese /w/). The small circles are the McCandliss training continua linearly interpolated between the anchors. The anchor stimuli are nested circles, and the midpoints are larger filled circles.

egories are nested within an existing category. A second approach, advocated by McClelland, McNaughton, and O'Reilly (1995), is to assume that there are two complementary learning systems in the brain: a hippocampal one that quickly learns conjunctions of features and a cortical one that gradually discovers higher order statistical structure in the hippocampal memories and integrates them with prior memories. This system allows stable long-term knowledge as well as the rapid learning of new categories that overlap with the existing categories.

We suggest that a similar complementary system may underlie the language learning in the McCandliss et al. (2002) experiment. Specifically, we assumed that (1) the training results in new auditory categories that interact with the representation layer without directly modifying the Japanese attractor structure, and (2) the new categories are mediated by a pool of units that is functionally separate from the L3 pool that mediates the Japanese attractor structure. In our approach, the R and L response units are placed in a new pool that has within-layer inhibition and bidirectional connections to L2 (we shall refer to the preexisting pool of L3 units as the *slow learning pool*, or SLP, and the new pool as the *fast learning pool*, or FLP). The learning rate between L2 and the FLP is much higher than that between L2 and the SLP, which ensures that the rapid intra-experiment learning does not catastrophically modify the preexisting tendencies in the network. A further advantage to this organization is that the FLP units can register the network's "R" and "L" labeling responses and receive feedback when it is available.

Incorporating outcome information into Hebbian learning. A central goal of our work was to explore the possibility of relying on Hebbian learning to account for the learning advantage of feedback. Although the feedback in the McCandliss et al. (2002) experiment indicated only whether the response was correct, subjects could easily use this information to determine which alternative was correct, since there are only two alternatives. We therefore assumed that the outcome information results in external input to the units representing the alternatives, with excitatory input being provided to the correct response unit and inhibitory input to the incorrect alternative (this is referred to as *soft-clamping*).[2] We also assumed that the size of these inputs was moderate, so that the units' final activation levels reflected a combination of the outcome-based input and the result of processing the current stimulus.

Modulation of the learning rate by the "confidence" of response unit activation. One of the key aspects of the McCandliss et al. (2002) experiment was the very slow learning in the fixed-without-feedback condition. Capturing this aspect of the data proved to be quite challenging, since competitive Hebbian learning tends to lead to one of two outcomes, neither of which match the pattern seen in the behavioral data: (1) If neither unit has a substantial initial advantage, one tends to win for one of the stimuli, whereas the other tends to win for the other; Hebbian learning strengthens this contrastive tendency, and the network rapidly learns to distinguish the two stimuli. (2) If initially one response unit has a substantial advantage over the other, that unit tends to win the competition between the response units for stimuli from both categories. Hebbian learning then strengthens the weights to this unit, allowing it to win even more often until it very quickly takes over the stimulus space for both categories. In short, competitive Hebbian learning is very rapid for appropriate and inappropriate learning outcomes; in the latter cases, the rapidity is counterproductive since it is very difficult for the system to recover from the inappropriate state. We therefore considered the possibility that a Hebbian system might trade off the overall rapidity of learning in exchange for a greater chance of appropriate learning. In particular, we posited that the rate of learning is proportional to the "confidence" of the network in its response (defined as the absolute difference between the activations of the R and L units). Under many conditions, the likelihood of being correct will tend to be lower when the confidence is lower, and in such cases, the "confidence modulation" would prevent the network from committing itself to an inappropriate outcome.

This modulation has important consequences for the current model. At the start of training, confidence is low for both correct and incorrect responses. In the feedback conditions, the soft-clamped outcome information enhances confidence for correct responses and lowers it for incorrect ones, allowing fairly rapid learning. In the fixed-without-feedback condition, confidence is low for both correct and incorrect responses, resulting in very slow learning that prevents a single response category from taking over the stimulus space for both categories. In the adaptive-without-feedback condition, the confidence tends to be higher on average (because the stimuli are exaggerated and easier to distinguish); therefore, learning progresses more quickly than in the fixed training condition.

Phase 1: Acquisition of Japanese /ɾ/ and /ɰ/

Architecture. The network consists of three layers (Figure 4A). The first two layers (L1 and L2) are both 20×20 grids, and L1 is topographically mapped to L2 such that the weight between L2 unit (a, b) and L1 unit (c, d) is $\exp(-[(a - c)^2 + (b - d)^2] / \beta_{21})$, with $\beta_{21} = 15$. The incoming L1 weight vector to each L2 unit is normalized to have unit magnitude, and the input space does not wrap around. The third layer consists of a single pool of units with 5 units (henceforth, the SLP). The SLP is bidirectionally connected to L2, with L2→SLP weights initialized from *Uniform*(0, 0.06) and SLP→L2 weights from *Uniform*(0, 0.0005). Each L2 unit is connected to itself with a weight of +1.0 and to all others in its layer with a weight of −0.07; each SLP unit is connected to itself with a weight of +2.0 and to all others in its pool with a weight of −2.0.

Network dynamics and weight update. The activity update was exactly the same as that in Simulation 1: $dt = 0.2$ for 30 timesteps, $\sigma_{\text{input}} = 0.02$, $\sigma_{\text{net}} = 0.2$, $gain_{\text{act}} = 0.5$, and $wscale_{kj} = 1$ for L1→L2 and L2→L3 weights and 5 for L3→L2 weights. The maximum weight vector magnitude was 1.0 for all projections, and there was no learning on L1→L2 weights. The learning rate on each trial was modulated by the model's "confidence" in its learning, defined as the absolute difference in activity (after settling) between the winning unit and the next most-active unit (Usher & McClelland, 2001). The learning rate η for the trial was $\eta = \eta_{\max} \cdot [1 - \exp(-confidence \cdot \beta_{\text{conf}})]$, where $\eta_{\max} = 0.004$ and $\beta_{\text{conf}} = 10$.

Inputs. The input space to the network was the two-dimensional space of $F2$ and $F3$ onsets, transformed from Hertz to Barks (the Bark scale resembles a log transformation of the formants and is an approximation of initial auditory processing, Kewley-Port & Atal, 1989).[3] The limits of the input space were defined as $F2_{\min} = 450$ Hz, $F2_{\max} = 2400$ Hz, $F3_{\min} = 1050$ Hz, and $F3_{\max} = 5500$ Hz (Figure 4B). For each input stimulus, the $F2$ and $F3$ onsets were converted into a location in the input grid:

$$x' = (N_x - 1) \cdot [(f2 - F2_{\min})/(F2_{\max} - F2_{\min})] + 1$$
$$y' = (N_y - 1) \cdot [(f3 - F3_{\min})/(F3_{\max} - F3_{\min})] + 1,$$

where $N_x = 20$ and $N_y = 20$. The activity of the input unit (a, b) was defined as $\exp(-[(x'-a)^2 + (y'-b)^2] / \beta_{\text{input}})$, with $\beta_{\text{input}} = 9$. The resulting input pattern was normalized to have a magnitude of 1.0. Exemplars of the two Japanese categories were drawn from bivariate Gaussian distributions. The mean $F2$ and $F3$ onset values, derived from Japanese speakers' utterances of /ɾa/ and /ɰa/ (data from Guion et al., 2000),[4] were 1340 and 2400 Hz for /ɾ/ and 625 and 2700 Hz for /ɰ/. Following our hypothesis about the formation of a large Japanese /ɾ/ attractor (and generally consistent with the data reported in Lotto, Sato, & Diehl, 2004), we assumed that /ɾ/ has a fairly large variability, with $F2$ $SD = \sqrt{6}$ Barks and $F3$ $SD = \sqrt{5}$ Barks. For /ɰ/, we used $F2$ $SD = \sqrt{0.15}$ Barks and $F3$ $SD = \sqrt{0.15}$ Barks. Finally, 1,600 exemplars were drawn from /ɾ/ and 400 were drawn from /ɰ/.

Training. The input stimuli were presented to the network 8,000 times in random order (without supervision), and the weights were updated after the network had settled in response to each input.

Phase 2: R/L Training

Architecture. In Phase 2, a second pool with two units (the FLP) is introduced in the third layer. This pool is bidirectionally connected to L2 and has the same internal connectivity as the SLP. Both the L2→FLP and FLP→L2 weights are initialized from *Uniform*(0, 0.0005). There is no interaction between the SLP and the FLP.

Dynamics and weight update. The activity update was exactly the same as that in Phase 1. The learning rate η on each trial was modulated by the confidence, with $\eta_{\max} = 0.001$. The "confidence" on a given trial was defined to be the absolute difference in activity—after settling—between the L unit and the R unit. Once the confidence was calculated, the L2↔FLP weights were updated with learning rate η and L2↔SLP weights with learning rate $\eta/20$.

Inputs. McCandliss et al. (2002) used two stimulus continua (*rock–lock* and *load–road*). We started with the $F2$ and $F3$ onsets of the natural utterances (the "anchor" stimuli).[5] If A_L and A_R are the $F2 \times F3$ locations of the /l/ and /r/ anchors, then the location A_k of stimulus index k was calculated as $A_L + k \cdot (A_R - A_L)$. The stimulus indices ranged from −0.6 to +1.6, in steps of 0.05. Once these locations were calculated, the corresponding input patterns were generated as in Phase 1. McCandliss et al. also established *midpoints* for the two continua, defined as the location of the crossover from <50% to >50% /r/ responses by native English speakers. These locations were 0.45 for *lock–rock* and 0.35 for *load–road*, and we used the same values for our continua also.

R/L pretraining. The McCandliss training assumed that Japanese listeners are initially able to distinguish extreme /r/ and /l/. In order to set up these initial response preferences and to match the pretraining experimental data, the L2→FLP weights were set to be initially quite weak. Next, one unit in the FLP was designated as the R-response unit and the other as the L-response unit. Then, the most exaggerated stimuli (stimulus indices −0.6 and +1.6) were presented to the network 300 times, and soft-clamped feedback was provided (see the *Training conditions* section). Confidence modulation was not used during this pretraining (i.e., $\eta = \eta_{\max} = 0.001$).

Training conditions. The network trained in Phase 1 was the starting point, serving as a generic "Japanese listener." One hundred twenty-eight instances of this network (32 in each training condition) were pretrained and then trained in a single session of 3,000 trials. Half the networks were trained on *rock–lock* and the other half on *road–load*. In all conditions, the labeling response of the network was taken to be the FLP unit with the greater activation after the final time step. In the no-feedback conditions, the weights were updated using the activity values at the final time step. In the feedback conditions, the net input of the correct FLP unit was incremented by 0.2 after the final time step, and the net input of the incorrect FLP unit was decremented by 0.2 (with a net input floor at 0.0). Then, the output activations were recalculated, the confidence was recomputed, and the weights were updated.

In the adaptive-stimulus conditions, the stimuli were paired into 37 "levels." If the midpoint of the continuum was designated m, then Level 0 consisted of the pair $(m - 0.05, m + 0.05)$, Level 1 had $(m - 0.1, m + 0.05)$, Level 2 had $(m - 0.1, m + 0.1)$, and so on. The adaptive training began at Level 30. When the network made eight correct responses in a row, the level was decreased by 1 (with a floor at Level 15); if the network made an error, the level was increased by 1 (with a ceiling at Level 37). In the fixed-stimulus conditions, the stimulus pair at Level 15 was used as the training stimuli (henceforth, the "fixed stimuli").

There are two differences between the above protocols and the actual experiment. In the experiment, the fixed stimuli were at Level 8 instead of Level 15, and there was no floor in the adaptive training. These restrictions are necessary in the model because if the stimuli are very close to each other, there is (1) greater competition between the L3 units, which (2) lowers the overall activity level and the amount of recurrent excitation, thereby (3) decreasing the discriminability peak at the category boundary. The use of Level 15 brought our model closer to the data regarding the starting place of learning and allowed us to

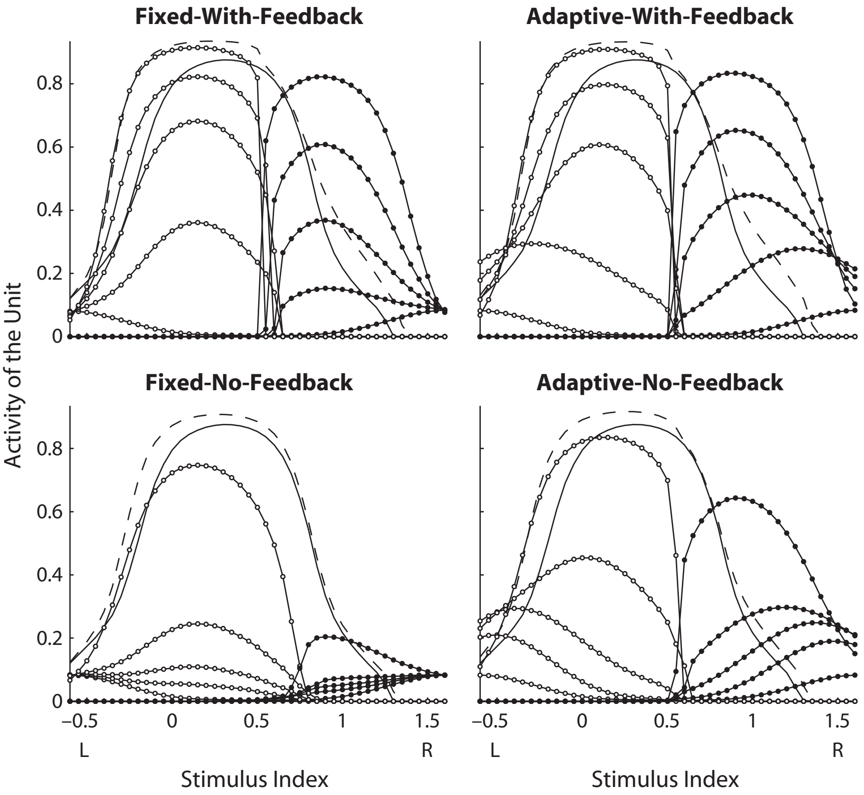

Figure 5. The activity of the FLP units for stimuli on the *lock–rock* continuum at four different stages of the training (immediately after pretraining and after 1,000, 1,500, 2,000, and 3,000 updates), showing the response of the L unit (empty circles), the R unit (filled circles), the Japanese tap unit before R/L training (unmarked solid line), and after R/L training (dashed line).

focus on the overall trend in the data (viz. the presence of a discriminability peak).

Results and Discussion

The learning of the network is best seen in the response of the L and R units to different stimuli, with both the input and integration noise turned off. Figure 5 shows these responses for the four training conditions at different stages of learning. The figure also shows the activity of the SLP unit that is associated with the Japanese tap attractor (effectively, the strength of the Japanese tap attractor). Crucially, the L-side stimuli are more prone to fall into the Japanese tap attractor than are the R-side stimuli, an asymmetry that follows directly from the layout of the input space. An L-side stimulus, therefore, activates the tap attractor, which recurrently excites L2, which increases L2 activity levels, which causes faster Hebbian learning. Thus, the L-unit in the FLP tends to (1) learn faster than the R-unit, (2) develop the same mapping as the Japanese tap unit, and (3) take over parts of the stimulus space corresponding to the English /r/, causing an overall bias toward L-labeling. A second point about Figure 5 is that slight learning also occurs for the Japanese tap unit in the SLP. The learning is subtle because of the slow learning rate; however, one can see that the attractor increases slightly to encompass both the /l/ and /r/ stimuli (Flege, 1995).

Categorization

Figure 6 shows the categorization curves during the training for the subjects and the model, and Figure 7 shows the time courses of the learning. The subjects' time course was evaluated by periodically presenting probe stimuli; every 20 trials, one of the fixed stimuli was randomly chosen and presented, with feedback being provided only in the feedback conditions. The networks' learning was evaluated in the same manner. The model's learning is a little faster than that of the subjects, but in general, the network captures the relative rates of learning: Learning is fastest in the fixed-with-feedback condition, followed by adaptive-with-feedback, adaptive-without-feedback, and finally, fixed-without-feedback. From Figure 6, the experimental training duration of 1,500 trials appears to correspond to approximately 1,500 weight updates in the model. Moreover, the model captures the interesting pattern that although posttraining categorization was better for the adaptive than for the fixed condition, in the no-feedback condition, the time courses of learning were very similar. Finally, the categorization curves (Figure 6A) show the bias toward L-labeling.

Two discrepancies between the experiment and the model may be noted. The first concerns the effect of extended training. In the experiment, the extended training improved the categorization in the fixed-without-feedback condition on the /l/ side of the continuum. The model also shows improvement, but the categorization curve as a whole is more biased toward L-labeling. However, this discrepancy may be incidental. While the /l/ bias of the categorization is not prominent in the McCandliss et al. (2002) data, we have observed it in ongoing experiments that use *lock–rock* and synthetic *la–ra* continua, sometimes to the extent that subjects classify all the stimuli as /l/. The second discrepancy concerns the time course of learning in the fixed-with-feedback condition. The experimental data (Figure 7B) show a very rapid initial increase (much faster than the adaptive-with-feedback condition), and the model does not. We shall return to this point in the discussion when assessing the effectiveness of Hebbian learning.

Transfer

Figure 8 shows how training on the *lock–rock* or the *load–road* continuum transfers to the other one (the categorization curves are averaged over both directions of transfer). The model's transfer is quite good by 2,000 weight updates and somewhat better than the experimental transfer curve after 1,500 training trials. The superior performance by the model is most likely because of its restriction to the $F2 \times F3$ input space. The actual sounds differ along dimensions other than $F2$ and $F3$, such as the overall intensity contour and the vowel context (/o/ vs. /ɑ/), and these additional differences might impair transfer in the actual experiment. However, there are two points that suggest that the $F2 \times F3$ space is a useful simplification. For both model and data, the transfer curves are (1) shifted toward the /r/ side of the continuum and (2) are quite poor for /l/ stimuli in the adaptive-without-feedback condition. In the model, the former is due to the influence of the Japanese tap attractor over the $F2 \times F3$ space, and the

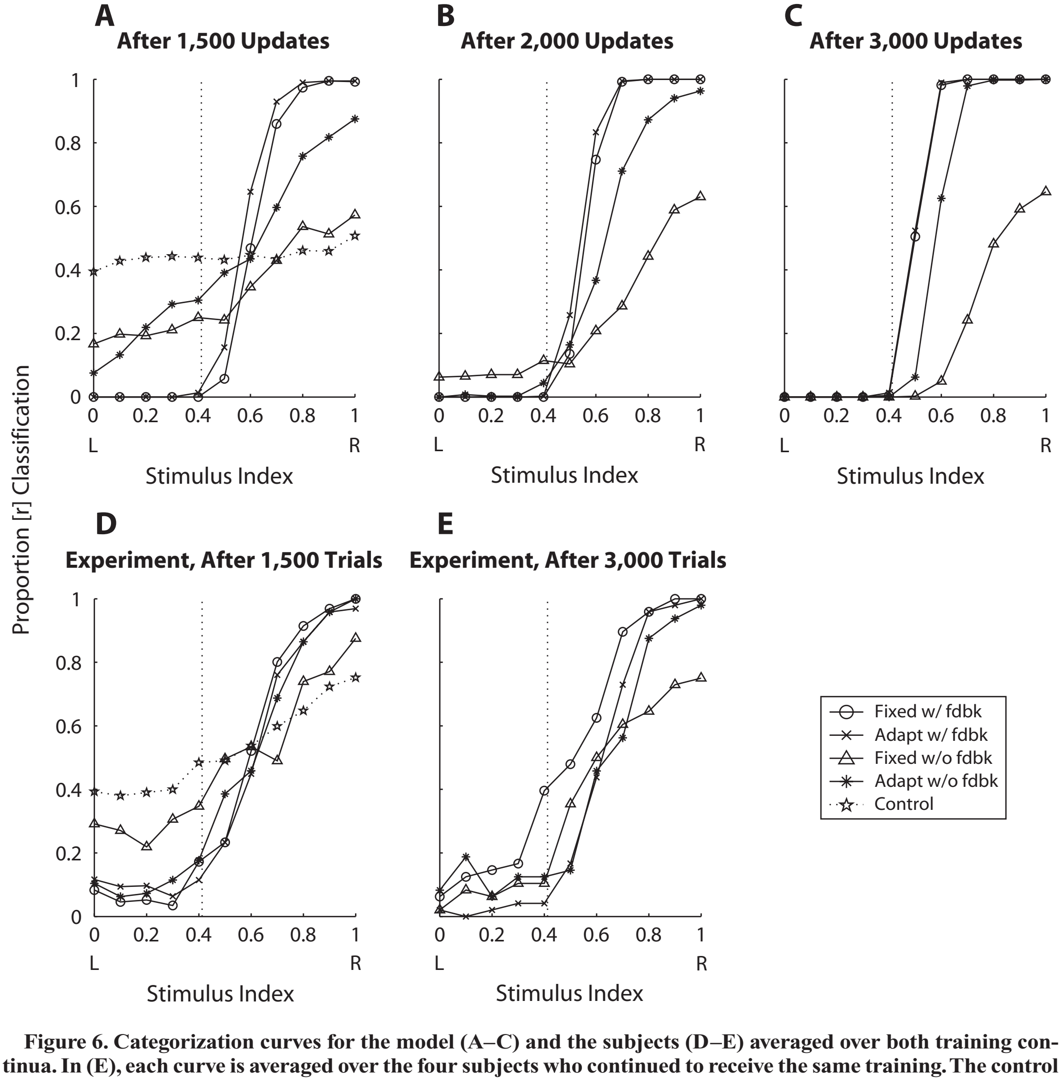

Figure 6. Categorization curves for the model (A–C) and the subjects (D–E) averaged over both training continua. In (E), each curve is averaged over the four subjects who continued to receive the same training. The control data are the average categorization curves prior to any training. In the fixed conditions, the midpoint between the fixed stimuli is 0.475 for *lock–rock* and 0.375 for *load–road*. In the adaptive conditions, the midpoints are 0.45 and 0.35, respectively. The vertical dotted line shows (0.35 + 0.375 + 0.45 + 0.475)/4 = 0.4125.

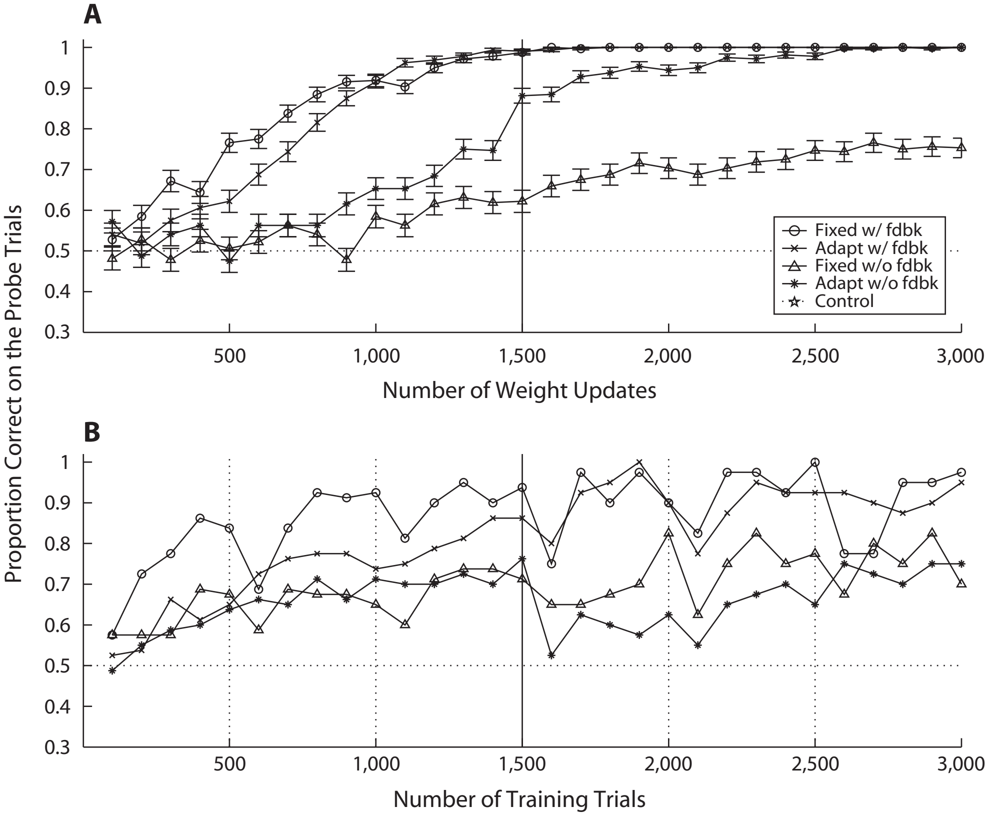

Figure 7. The time course of learning for (A) the model. The standard error bars are calculated as $\sqrt{[p(1 - p) / N]}$, where p is the proportion correct and N is the number of data points included in the average. There are 32 instances per condition and each data point is averaged over 10 probe values, so $N = 320$. (B) The subject data. The vertical dashed lines mark the end of a single session. For trials 1,500–3,000, the curves are averaged over the four subjects who continued to receive the same training.

latter is due to the layout of stimuli in the $F2 \times F3$ input space (Figure 4B). Specifically, the *rock* and *road* anchors are closer to each other than are the *lock* and *load* anchors; consequently, it is easier to transfer /r/ across continua than /l/. These similarities suggest that stimulus relationships in the $F2 \times F3$ space capture some of the acoustic relationships between the actual stimulus sounds.

Discrimination

The discrimination ability of the model was evaluated in the same manner as that of the subjects. In the *slide* test, eight pairs of stimuli were presented for same–different discrimination. The distance between the stimuli in each pair was fixed at 0.3, and the midpoint of the pair was gradually increased (0.15 to 0.85 in steps of 0.1). In the *expand* test, the midpoint of each stimulus pair was fixed at the midpoint of the continuum, and the distance between the stimuli in each pair was gradually increased (0.1 to 1.5 in steps of 0.2).

These two tasks were mapped onto the network in the following manner. Each stimulus in a pair was presented to the network, and once the network settled, the activity of the representation layer was recorded (the weights were not updated during the tests). The patterns of activity of the two stimuli (their "perceptual representations") were then used to generate a same–different response. We used a variant of the "differencing" strategy (Macmillan & Creelman, 1991): First, we presented the "same" and "different" stimulus pairs to the network and calculated the Euclidean distance between each pair of perceptual representations. These distances were then averaged to get a mean "perceptual distance" for each unique pair. Next, we estimated a mapping from the perceptual distance to the probability of a "different" response.[6] Figures 9 and 10 directly report this probability.

Figure 9 shows the results of the slide test for the model and the subjects. The control subjects show a gradual increase in the discriminability toward the /r/ end of the continuum. After the R/L training, the subjects in the adaptive and the fixed-with-feedback conditions develop a peak at the category boundary that rides on top of the existing slope. The model generates qualitatively similar discrimination curves. The pretraining shape of the curve occurs because the /r/ end of the continuum is closer to the boundary of the Japanese tap attractor (see Figure 4B). With training, the R and L units develop two subsidiary attractors within the larger, shallower basin of the Japanese tap attractor, thereby creating a discrimination peak at the boundary (the location of the peak coincides with the crossover point in the categorization curve). The peak development is mediated by the

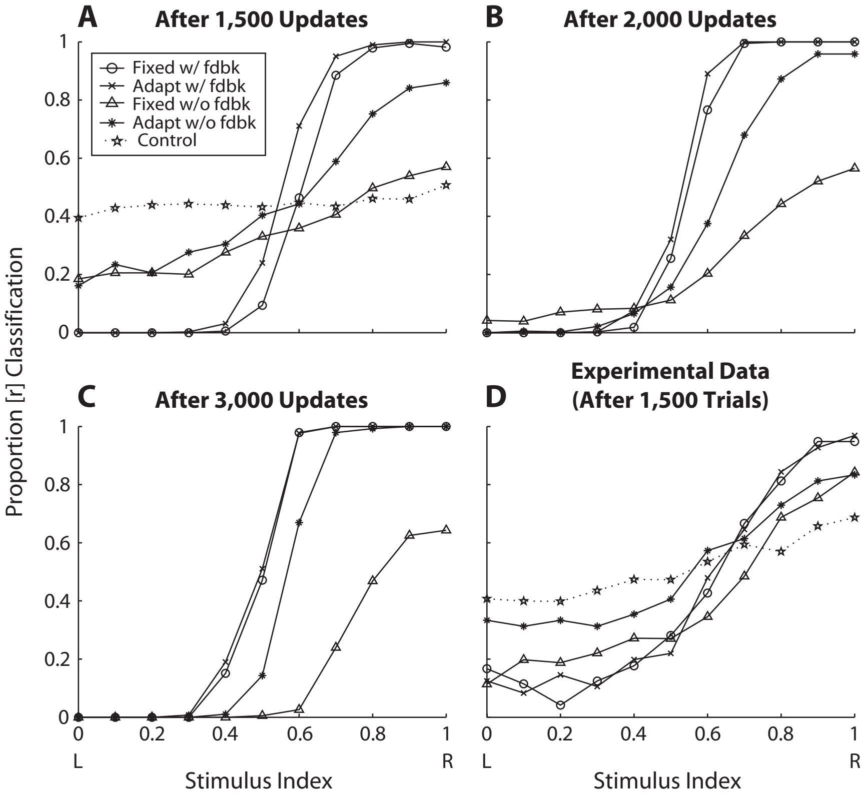

Figure 8. The performance of the model (A–C) and the subjects (D) on the transfer continuum (averaged over both *lock–rock* and *load–road* transfers).

strong lateral inhibition in FLP; this ensures a sharp change from one category to the other and maximizes the perceptual distance between stimuli straddling the boundary.

The results of the expand test are shown in Figure 10. The experimental data are rather noisy, but generally there is an increase in discriminability after training, with the largest change occurring at a separation between 0.5 and 1.0. The model shows a similar increase. One difference is that the model's discrimination peaks at a separation just short of 1.0 and decreases thereafter. This is because each category is also a spatially bounded region in L2. As the stimuli move further away from categories, the attractors are only weakly activated; therefore, the perceptual representations are weaker. Some of the discrimination curves in the experiment also show a slight decrease, so there is tentative support that a spatially bounded representation may underlie the subjects' performance also.

One interesting relation between the data and the model concerns the timing between the development of the categorization and discrimination. In the data, the categorization for the feedback conditions is quite good after 1,500 trials, and the corresponding discrimination peaks are also well developed (Figures 6D and 9D). The categorization for the adaptive-without-feedback condition is also quite good, but its discrimination peak appears to be less developed. This suggests that the discrimination peak need not always "keep pace" with the categorization, but can lag behind to some extent. The model exhibits an intriguingly similar pattern of lag. The categorization for the feedback conditions is quite good after 1,500 weight updates, and the discrimination peaks take 500 more updates to develop. The categorization for the adaptive-without-feedback condition is quite good after 2,000 updates, and its peak takes 1,000 more updates to develop. Thus, the model exhibits an exaggerated version of the lag observed in the data, with the amount of lag being smaller in the more effective training conditions. The lagging is due to the development of the attractor dynamics. First, the categorization is stabilized, then (after more training) the attractors get strengthened and affect the discrimination. The lag is greater with adaptive-without-feedback mainly because the learning is slower; thus, the attractors take longer to strengthen.

Individual Differences

As was noted earlier, there were individual differences in the no-feedback training, with some subjects in the fixed condition learning as quickly as "good" learners in the adaptive condition and vice versa. Some of these differences may be due to variations in the subjects' perceptual spaces. For example, the Japanese tap attractor may be much more powerful or cover a larger extent, causing the training stimuli to be perceptually much more similar to each other and requiring many more training trials before the /r/ and /l/ categories get established.

The model as was previously described assumes that all "subjects" (i.e., instances of the network) have the same initial state. We attempted to simulate variations in the initial state by manipulating the relation between the initial state and the training stimuli. Specifically, we assumed that enlarging the extent of an attractor is equivalent to shrinking the length of a continuum that cuts across that attractor.[7] The shrinking was implemented as follows: Let $s_1 \ldots s_{45}$ be formants of the stimuli, and let s_m be the midpoint of the continuum. Then, the rescaled stimuli $rs_1 \ldots rs_{45}$ were defined by $rs_i = stimscale \cdot (s_i - s_\mathrm{m}) + s_\mathrm{m}$. We then reran the simulations, with *stimscale* ranging from 0.3 to 1.0. For each value of *stimscale*, 16 instances of the network were trained on the *lock–rock* continuum (four in each of the training conditions) for 3,000 weight updates. All the instances were pretrained with the extreme stimuli of the unmodified continuum (i.e., *stimscale* = 1.0), so the rescaled continuum was only used during the actual training. All other parameters were the same as those in the previous simulation.

Figure 11 shows the results from the simulation in terms of the slope and bias in the categorization curves, with greater slope and smaller bias indicating better categorization. At *stimscale* = 1.0, we see the expected order of categorization ability—the two feedback conditions, then the adaptive-without-feedback condition, and finally, the fixed-without-feedback condition (the latter with a strong L-bias). With decreasing *stimscale*, the categorization becomes more difficult, indicated by a decreasing slope and (because of increasing dominance of the Japanese tap attractor) an increasing labeling bias. Each individual subject in the experiment may be thought of as being at a different location on the *stimscale* continuum: A subject may be unable to learn in the fixed-without-feedback condition but thrive in the others (e.g., at *stimscale* = 0.7 or 0.8). Conversely, a subject in the adaptive-without-feedback condition may fare as badly as a subject in the fixed-without-feedback condition (e.g., if the former has *stimscale* = 0.6, and the latter has *stimscale* = 0.9). Of course, this explanation is only one of several possible accounts for intersubject variation, but in the discussion, we shall consider ways to diagnose the "stimscale–attractor dominance" of a subject before training.

The curves in Figure 11 can also be interpreted as indicating the robustness of the training conditions. The fixed-with-feedback condition collapses at *stimscale* = 0.6, whereas the adaptive conditions show more "graceful degradation." The brittleness of the fixed-with-feedback condition (it either works really well or not at all) stems from the use of soft-clamped supervision. If hard-clamped supervision or much stronger soft clamping were used instead (or equivalently, the "strength" of the feedback to the subject was varied in some way), then learning would be possible for smaller *stimscale* values.

GENERAL DISCUSSION

The Hebbian attractor model elucidated in Simulations 1 and 2 is able to capture several key phenomena in speech

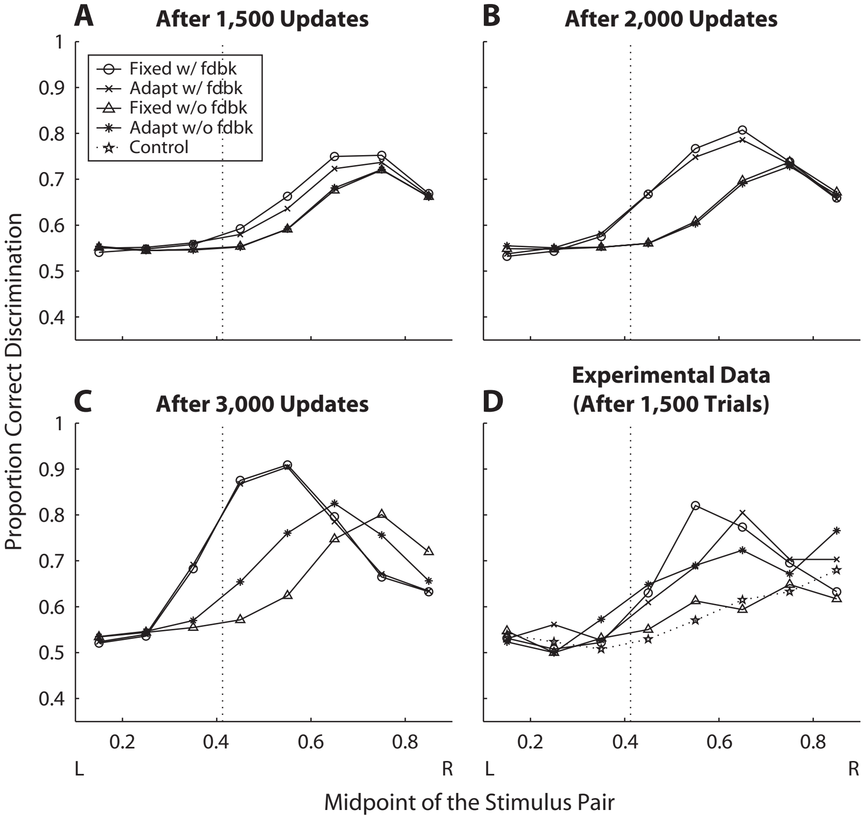

Figure 9. The results of the slide test for (A–C) the model and (D) the subject data. The vertical dotted lines mark 0.4125 (see the caption for Figure 6).

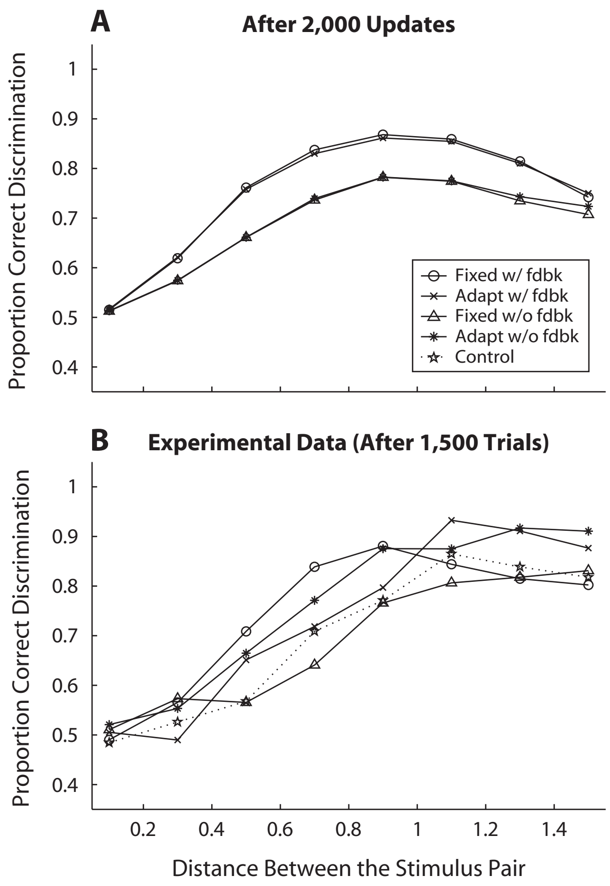

Figure 10. The results of the expand test for (A) the model and (B) the subject data.

perception, such as unsupervised category acquisition, graded category membership, and a systematic relation between category membership and discriminability. In addition, the model is able to capture many of the results from McCandliss et al. (2002). These include (1) the effectiveness of adaptive-without-feedback training, (2) the relative time course of learning in different feedback and stimulus conditions, (3) a strong bias toward L-labeling in posttraining categorization, (4) transfer of training to a new stimulus continuum, and (5) an acquired increase of discriminability at the category boundary. Furthermore, the model provides a possible basis for some aspects of the individual differences seen in learning.

More generally, our model gives a mechanistic account of how perception is influenced by prior experience, how difficulty in discrimination may be related to difficulty in second language learning, and how outcome information and exaggerated input may be used to facilitate learning. The account, similar to that articulated by Flege (1995) and Kuhl (2000), hypothesizes that initial experience influences perception through the development of dynamical attractors that are automatically engaged by the current stimulus. In adulthood, learning occurs after a perceptual representation is modified by the attractors; therefore, the learning is slower and also tends to further entrench the attractors. Exaggerated input allows the system to "break free" of the attractors and establish distinct representations, thereby promoting learning even when feedback is not provided. The facilitation of learning by exaggerated inputs may also underlie other phenomena in speech learning. For example, 6–12-month-old Mandarin infants who are exposed to clearer (i.e., more exaggerated) vowel sounds in infant-directed speech are better at discriminating those vowel sounds (Liu, Kuhl, & Tsao, 2003), and in fact, English and Japanese mothers tend to exaggerate the durations and formants of infant-directed speech in language-specific ways (Werker et al., 2007; see also Kuhl et al., 1997).

Assessment of the Principles Underlying the Model

As was noted earlier, our goal for this model was to examine whether four key principles—interactive competition leading to attractors, increased attractor strength with learning, Hebbian learning, and topographic maps—may be sufficient to account for the perceptual learning in the McCandliss et al. (2002) experiment. What conclusions can be reached on this score? Our short answer would be that the principles, augmented with the additional principle of confidence modulation of the learning rate, have brought us fairly close to a complete account of the data. We first discuss how the principles contributed to the successes of the model and then consider some open issues.

The attractor dynamics and the increased attractor strength with learning are responsible for the L-bias in the categorization curves, the pretraining difficulty in discriminating /l/ and /r/ stimuli, the development of a discrimination peak at the category boundary, and the lag between the improvement of the categorization and the development of the discrimination peak. More generally, these principles help explain why nonnative contrasts that map onto a single native category are especially difficult to tell apart (consistent with the predictions made by the perceptual assimilation model; Best, 1995) and are especially difficult to learn (consistent with the speech learning model; Flege, 1995). Furthermore, the principles predict that continual exposure to unexaggerated English /r/ and /l/ sounds may not be helpful for Japanese listeners: An attractor that inadvertently pulls in both /r/ and /l/ only gets strengthened by further exposure to the sounds.

The Hebbian learning promotes the symmetric connectivity that is necessary for stable attractor dynamics (Grossberg, 1976, 1988). Moreover, it allows a uniform account of both unsupervised and supervised learning. Essentially, Hebbian learning reinforces the pattern of activity in place during the weight update; consequently, any mechanism that promotes the formation of correct activity patterns will result in successful learning. When outcome information is present, the correct activity pattern is promoted through soft clamping. When outcome information is absent, the correct activity patterns may be promoted using initially exaggerated inputs. The confidence-modulation of the learning rate aids in the learning by ensuring that the network does not prematurely commit to a learning path. The idea that something like confidence might be used to modulate processing is

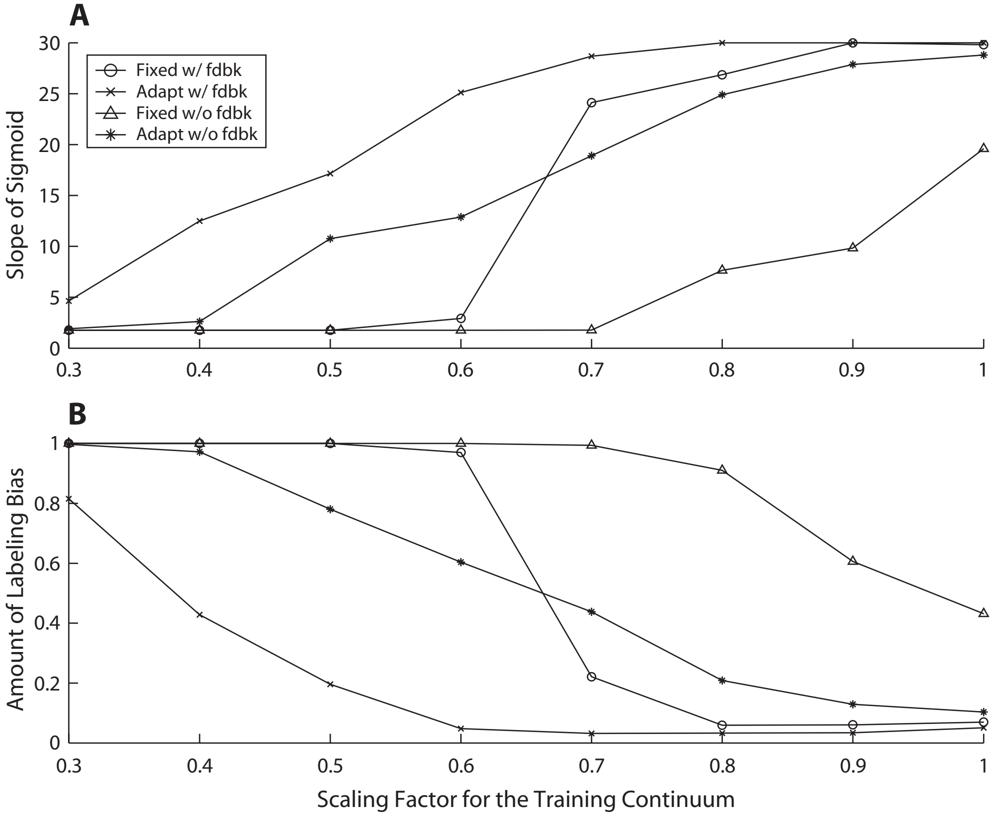

Figure 11. The effect of scaling down the continuum on categorization ability after 3,000 weight updates. (A) The slope of the categorization curve, measured by the mean gain of sigmoids fitted to the categorization curves. Higher values indicate more categorical classification. (B) The amount of labeling bias. Let *m* be the area above or below the categorization curve (whichever is greater), scaled to be between 0.5 and 1.0. Then, labeling bias ≡ 1 − 2 <*m*>, where <*m*> is the mean area. Biases near 0.0 indicate a symmetric categorization curve; biases near 1.0 indicate that all the stimuli on the continuum were given the same label.

strongly promoted in other work. Cohen, Aston-Jones, and Gilzenrat (2004) describe a connectionist architecture for attentional control in which the "conflict" (the competition between alternative responses) is monitored and used to regulate allocation of attention. The conflict signal is quite similar to our "confidence" signal and could be used to modulate learning rate. This could occur, for example, by modulating activation in dopamine neurons, which may in turn regulate learning rate at synapses undergoing modification (Montague, Hyman, & Cohen, 2004).

Finally, the topographic representation allows a simple explanation for learning with adaptive stimuli and for the transfer effects, since proximity in stimulus space maps directly into overlap between the corresponding perceptual representations. We do not suggest that topographic representations are necessary for speech perceptual learning. However, their utility in this model, together with neurophysiological evidence for topographic organization of complex visual features (see, e.g., Tanaka, 1997), suggest that speech learning and processing are also served by topographic representation of phonetically relevant parameters.

Although the principles are successful in accounting for many aspects of the data, two issues remain. First, we had to stipulate a minimum distance between the adaptive stimuli in the model (i.e., introduce a floor at Level 15), a restriction that was absent in the experiment. Second, the advantage of fixed over adaptive training when feedback is provided is weaker in the model than in the data (Figures 7A and 7B). We now discuss the difficulties in more detail and evaluate their implications for the principles.

As was noted earlier, the stipulation of a minimum distance between the adaptive stimuli is necessary because greater proximity between stimuli leads to greater competition in L3, which in turn lowers the activity of the R and L units. Essentially, the competition in L3 does not lead to a binary state (where the winning and losing units settle at 1.0 and 0.0, respectively). The competition can be made more binary by increasing the self-excitation and lateral inhibition. Doing so, however, tends to disrupt learning because the network is always fully confident in its responses, even when it makes erroneous responses early in training, thereby prematurely committing itself to inappropriate learning outcomes. However, this shortcoming may be overcome by a mechanism recently implicated in real neural systems. In our model, the within-layer weights are fixed; therefore, the characteristics of the competition do not change over training. Some have suggested, however, that inhibitory connection weights may be up-regulated as excitatory input increases (Foeller & Feldman, 2004; Lamsa, Heeroma, & Kullmann, 2005). Such a mechanism could allow the inhibitory weights to change, such that the inhibition is moderate early in training (resulting in graded FLP activity, low confidence, and relatively slow learning) and high later in training (resulting in binary FLP activity, high confidence, and relatively fast learning). With such an extension to the within-FLP dynamics, the network may be able to successfully learn even with closely spaced inputs, thereby eliminating the need for a minimum distance between adaptive stimuli.[8]

The second issue (that the advantage of fixed training when feedback is provided is weaker in the model than in the data) is more involved because of a complexity in the data. The subjects' performance in the fixed-with-feedback condition at the beginning of each session is almost at the level of the adaptive-with-feedback condition, jumping up within each session and then falling back again at the beginning of the next (Figure 7B). If we take the performance at the beginning of each session as an indicator of a longer-lasting component of learning, the model would seem to capture this component quite well. Of course, the basis of the short-lived within-session advantage still remains to be specified. We leave for future research to consider whether a distinction between short- and long-lasting components is useful and what the nature of the short-term component might be.

Comparison With the Model Proposed by Guenther and Colleagues

Guenther and his colleagues have taken an approach to auditory category learning that differs significantly from ours. The key idea behind their model (Guenther & Bohland, 2002; Guenther, Nieto-Castanon, Ghosh, & Tourville, 2004) is that the initial auditory layer has separate topographic projections to a category-learning layer (CL) and an auditory map (AM) that are responsible for categorization and discrimination judgments, respectively. The CL has inhibitory projections to AM, so a prototypical stimulus (1) causes greater CL activity, which (2) increases AM inhibition, which (3) decreases AM activity and makes it more susceptible to noise, thereby (4) impairing discrimination judgments. Consequently, discriminability is worse for the prototypical stimulus than for nonprototypical stimuli.

Guenther et al.'s (2004) proposal differs from ours in several ways. One key difference is that they explicitly address different effects of categorization versus discrimination training (Guenther, Husain, Cohen, & Shinn-Cunningham, 1999). Thus far, we have only addressed categorization training and have not yet considered ways in which the model could be extended to address discrimination training. It will be interesting to explore whether an approach based on assigning every possible stimulus to its own distinct category can provide an account of performance in such discrimination tasks or whether another approach altogether will be required. A second difference between the models concerns their architectures. In our model, a more prototypical stimulus invokes a stronger attractor, thereby suppressing noise. Consequently, if an AX discrimination test is given after categorization training, then Prob("different" | same) should decrease toward the center of the category. In Guenther et al.'s (2004) model, a more prototypical stimulus suppresses stimulus-driven activity in the AM, thereby enhancing noise. Consequently, their model predicts that Prob("different" | same) should increase toward the center of the category.

The difference in architecture corresponds to another difference in predictions as well: Guenther et al. (2004) predicted that categorization training results in decreased AM activity for prototypical stimuli, whereas we assumed that the training results in increased L2 activity.[9] The data from Guenther et al. (2004) appear to indicate a decrease in activity, but there are two concerns. First, a general decrease in fMRI activity may mask a smaller area of increased, highly coherent neural activity (Recanzone et al., 1992a, p. 1080). Second, in other studies, R/L categorization training has produced widespread activity increases in the auditory cortex (Callan et al., 2004; Callan et al., 2003).[10] We therefore suggest that the existing evidence is ambiguous and that further research aimed at producing data that distinguish the two theories is warranted.

Directions for Further Tests and Extensions of the Model

One exciting direction for future elaboration of the model is the long-term consolidation of new second language learning. In our current implementation, the FLP allows fairly rapid learning that does not change the connections that have been built up in the SLP through years of experience. What happens once the subjects leave the experiment? One possibility is that the SLP may very gradually incorporate the newly acquired distinction. Following successful R/L training, the dynamics between the FLP and L2 add two new attractors to the system that correspond to English /r/ and /l/. If the training is sustained and intensive, the /r/ and /l/ attractors result in two consistently different patterns of L2 activity. Unused units in the SLP may become sensitive to one or both of these two patterns and slowly learn them. Thus, the FLP could act like "training wheels" for the SLP, stabilizing the L2 representations so that the slower pool can consolidate them more thoroughly. Exploratory simulations with the current model support this possibility and indicate that it is a promising direction for future modeling and experimental efforts.

Note that the FLP is not exclusive to adult second-language learning. Following the complementary memory systems hypothesis (McClelland et al., 1995), we suppose that the FLP exists side by side with the SLP throughout life. During first-language acquisition, the perceptual space is relatively uniform and easily shaped by the FLP activity; consequently, consolidation of FLP structure into the SLP would be rapid. Later in life, the perceptual space is inhomogenous and less amenable to shaping by the FLP; consequently, consolidation would be slower and more error prone.

Our model also allows other predictions about R/L training. If Japanese listeners are given a sliding-window discrimination test with stimulus pairs extending beyond the *lock–rock* and *load–road* anchors, then there should be a peak in the discriminability that indicates the boundary of the Japanese tap category (the location of the peak may vary among subjects). The further the peak is from the anchor stimuli, the greater the extent of the Japanese tap attractor and the greater the predicted difficulty of the subject in acquiring the R and L categories. Moreover, there should be a general L-labeling bias during and after training, with the amount of /l/ bias being inversely proportional to the amount of feedback (e.g., the percentage of trials on which feedback is given). The rationale is that learning without feedback is dominated by preexisting attractors; thus, new stimuli are more likely to be assimilated into existing categories. The feedback tends to counter this assimilation and allows the new category to establish itself (cf. the amount of L-bias in the different conditions in Figure 6).

Finally, although we have focused on speech and second-language acquisition, some of the phenomena are characteristic of perceptual learning in other modalities also. For example, Goldstone (1994) trained subjects to categorize patches of different sizes and brightness and found that subjects showed increased discriminability at category boundaries and instances of decreased discriminability for same-category items. This pattern may be explained with the current model if the two dimensions of the input layer are assumed to be size and brightness (cf. Goldstone, Steyvers, & Larimer, 1996). Similar networks may account for other instances of supervised visual category learning (Livingston, Andrews, & Harnad, 1998), unsupervised visual category learning (Rosenthal et al., 2001), and supervised auditory category learning (Guenther et al., 1999).

CONCLUSIONS

Interactive competitive networks with graded activation have been proposed for several perceptual and linguistic phenomena (McClelland, 1992; McClelland & Elman, 1986; McClelland & Rumelhart, 1981; Usher & McClelland, 2001) and share several properties in common with models addressing many other phenomena that have been proposed by Grossberg and collaborators (e.g., Carpenter & Grossberg, 1987). The results of the work reported here suggest that this framework can also be extended to several aspects of speech perceptual learning. In particular, our explorations indicate that four architectural principles—interactive competition leading to attractors, increased attractor strength with learning, Hebbian learning, and topographic maps—can provide a framework in which one can account for the basic pattern of success and failure in learning with and without outcome information. Instantiation of these principles in a concrete model also revealed the need for (1) an architectural differentiation into fast and slow learning systems, (2) uniform treatment of response predispositions and external outcome information in a manner consistent with Hebbian learning, and (3) modulation of the learning rate by the "confidence" of the network in its responses. The modeling also underscored the experimental data in need of clarification, such as the distinction between reinforcement and supervisory feedback and the relative advantage of fixed over adaptive training when feedback is provided.

Our effort has also revealed that there is far more work to do before we can claim to have a full understanding of speech perceptual learning. We have focused on acquiring a nonnative speech category distinction in adulthood, modeling data from a single published experiment. Several deficiencies in the model's account of these data remain to be fully addressed, and we have also indicated several possible extensions of the model to address issues beyond this single experiment. A fuller account is also needed of the initial development of spoken language perception. Whether the current approach can be extended to fully address this process as it occurs in the natural experience of young children in all language cultures is a question that can only be answered by future investigations.

AUTHOR NOTE

This research was supported by NIMH Grant MH64445 and Training Grant 5T32-MH019983-07. The authors thank Susan G. Guion for generously providing the recordings from her experiment and Mary L. Conway for her assistance with data analysis. Correspondence concerning this article should be addressed to J. L. McClelland, Department of Psychology, 450 Serra Mall, Jordan Hall, Bldg. 420, Stanford, CA 94305 (e-mail: jlm@psych.stanford.edu).

REFERENCES

Ahissar, E., Abeles, M., Ahissar, M., Haidarliu, S., & Vaadia, E. (1998). Hebbian-like functional plasticity in the auditory cortex of the behaving monkey. *Neuropharmacology*, **37**, 633-655.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, **84**, 413-451.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 167-200). Timonium, MD: York Press.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, **21**, 149-186.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage*, **22**, 1182-1194.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language contrast. *NeuroImage*, **19**, 113-124.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, & Image Processing*, **37**, 54-115.

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 71-90). New York: Guilford.

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics*, **62**, 843-867.

Edelman, S., & Intrator, N. (2002). Models of perceptual learning. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 337-354). Cambridge, MA: MIT Press.

Flege, J. E. (1992). Speech learning in a second language. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 565-604). Timonium, MD: York Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233-277). Timonium, MD: York Press.

Foeller, E., & Feldman, D. E. (2004). Synaptic basis for developmental plasticity in somatosensory cortex. *Current Opinion in Neurobiology*, **14**, 89-95.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, **123**, 178-200.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 243-248). Mahwah, NJ: Erlbaum.

Grajski, K., & Merzenich, M. M. (1990). Hebb-type dynamics is sufficient to account for the inverse magnification rule in cortical somatotopy. *Neural Computation*, **2**, 71-84.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**, 121-134.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, **1**, 17-61.

Guenther, F. H., & Bohland, J. W. (2002). Learning sound categories: A neural model and supporting experiments. *Acoustical Science & Technology*, **23**, 213-221.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, **106**, 2900-2912.

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., & Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, & Hearing Research*, **47**, 46-57.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America*, **107**, 2711-2724.

Herzog, M. H., & Fahle, M. (1998). Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics*, **78**, 107-117.

Hoshino, O. (2002). Dynamic interaction of attractors across multiple cortical networks as a neural basis for intersensory facilitation. *Connection Science*, **14**, 115-135.

Idiart, M., Berk, B., & Abbott, L. F. (1995). Reduced representation by neural networks with restricted receptive fields. *Neural Computation*, **7**, 507-517.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. *Journal of the Acoustical Society of America*, **118**, 3267-3278.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, **87**, B47-B57.

Jamieson, D. G., & Moroson, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, **43**, 88-96.

Kewley-Port, D., & Atal, B. (1989). Perceptual differences between vowels located in a limited phonetic space. *Journal of the Acoustical Society of America*, **85**, 1726-1740.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks*, **6**, 895-905.

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

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, **97**, 11850-11857.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, **277**, 684-686.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608.

Lamsa, K., Heeroma, J. H., & Kullmann, D. M. (2005). Hebbian LTP in feed-forward inhibitory interneurons and the temporal fidelity of input discrimination. *Nature Neuroscience*, **8**, 916-924.

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

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences*, **83**, 7508-7512.

Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, **6**, F1-F10.

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.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 732-753.

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.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). *Mapping the task for the second language learner: The case of the Japanese acquisition of /r/ and /l/*. Paper presented at the "From Sound to Sense: 50+ Years of Discoveries in Speech Communication" conference, Cambridge, MA.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, **2**, 89-108.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, **23**, 1-44.

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, and interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 655-688). Cambridge, MA: MIT Press.

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

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, **88**, 375-407.

McClelland, J. L., Thomas, A., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. In J. Reggia, E. Ruppin, & D. Glanzman (Eds.), *Disorders of brain, behavior and cognition: The neurocomputational perspective* (Progress in Brain Research, Vol. 121, pp. 75-80). Amsterdam: Elsevier.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, **50**, 271-285.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, **18**, 331-340.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioral control. *Nature*, **431**, 760-767.

Movellan, J. R., & McClelland, J. L. (2001). The Morton–Massaro law of information integration: Implications for models of perception. *Psychological Review*, **108**, 113-148.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Petrov, A., Dosher, B. A., & Liu, Z.-L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, **112**, 715-743.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.

Pleger, B., Foerster, A.-F., Ragert, P., Dinse, H. R., Schwenkreis, P., Malin, J.-P., et al. (2003). Functional imaging of perceptual learning in human primary and secondary somatosensory cortex. *Neuron*, **40**, 643-653.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, **55**, 899-910.

Recanzone, G. H., Merzenich, M. M., & Schreiner, C. E. (1992a). Changes in the distributed temporal response properties of SI cortical neurons reflect improvements in performance on a temporally based tactile discrimination task. *Journal of Neurophysiology*, **67**, 1071-1091.

Recanzone, G. H., Merzenich, M. M., & Schreiner, C. E. (1992b). Progressive improvement in discriminative abilities in adult owl monkeys performing a tactile frequency discrimination task. *Journal of Neurophysiology*, **67**, 1015-1030.

Rosenthal, O., Fusi, S., & Hochstein, S. (2001). Forming classes by stimulus frequency: Behavior and theory. *Proceedings of the National Academy of Sciences*, **98**, 4265-4270.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 7-57). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75-112.

Sirosh, J., & Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, **9**, 577-594.

Strange, W., & Dittmann, S. (1984). Effect of discrimination training on the perception of /r–l/ by Japanese adults learning English. *Perception & Psychophysics*, **36**, 131-145.

Sutton, G. G., Reggia, J. A., Armentrout, S. L., & D'Autrechy, C. L. (1994). Cortical map reorganization as a competitive process. *Neural Computation*, **6**, 1-13.

Syka, J. (2002). Plastic changes in the central auditory system after hearing loss, restoration of function, and during learning. *Physiological Reviews*, **82**, 601-636.

Takagi, N. (2002). The limits of training Japanese listeners to identify English /r/ and /l/: Eight case studies. *Journal of the Acoustical Society of America*, **111**, 2887-2896.

Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, **7**, 523-529.

Tarr, M. J., & Cheng, Y. D. (2003). Learning to see faces and objects. *Trends in Cognitive Sciences*, **7**, 23-30.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, **88**, 97-100.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550-592.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, **103**, 147-162.

White, S. (2001). Learning to communicate. *Current Opinion in Neurobiology*, **11**, 510-520.

Yamada, R. A., & Tohkura, Y. (1990). Perception and production of syllable-initial /r/ and /l/ by native speakers of Japanese. In *Proceedings of the 1990 International Conference on Spoken Language Processing* (pp. 757-760).

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., & Tohkura, Y. (2005). Effects of language experience: Neural commitment to language-specific auditory patterns. *NeuroImage*, **26**, 703-720.

NOTES

1. Recanzone et al. (1992a) reported the representations of trained and untrained skin only for the range of training frequencies (20–30 Hz). Hence, it is not clear whether the increased cortical activity was specific to location and frequency of stimulation or just the location alone. For current purposes, the key point is that topographic distribution of stimuli is one (though probably not the only) factor that results in increased activity and enlarged representations.

2. "Soft clamping" is typically used to influence the activities of some units while allowing them to participate in interactive processing. Presently, we use the term more loosely to refer to any adjustment of the net input.

3. We used the formula from Traunmüller (1990): $\text{Bark}(f) = [26.81 / (1 + 1960 / f)] - 0.53$.

4. The formant analysis of the Guion et al. (2000) sounds indicated the mean $F2$ onset for /ɰ/ to be approximately 900 Hz. In our preliminary simulations with this $F2$ location, /ɰ/ was too close to /ɾ/, and the /r/ and /l/ stimuli were often pulled into the /ɰ/ attractor. This behavior does not match experimental data, and in fact, the Japanese /ɰ/ category appears to assimilate more toward the American English /w/ (Yamada & Tohkura, 1990). Therefore, we shifted the mean $F2$ onset of /ɰ/ to 625 Hz, which is much closer to the American /w/.

5. The $F2$ and $F3$ onsets of the anchor stimuli were measured 70 msec from the start for *lock–rock* stimuli and 145 msec from the start for *load–road* stimuli (the onset locations are different because the *load–road* stimuli happened to have a longer initial pause). The formant values for the anchors were as follows: *lock* ($F2 = 1083$ Hz, $F3 = 2944$ Hz), *rock* ($F2 = 968$ Hz, $F3 = 1753$ Hz), *load* ($F2 = 1246$ Hz, $F3 = 2856$ Hz), *road* ($F2 = 939$ Hz, $F3 = 1732$ Hz).

6. The mapping was estimated as follows. (1) Assume there are N unique stimulus pairs in the slide test, so that the test yields N average perceptual distances for each condition in the model. We picked the condition with the largest range of distances; this was the fixed-with-feedback condition after 2,000 updates. (2) From the experimental data, we calculated the average probability of a "different" response for each of the N unique stimulus pairs in the slide test for the fixed-with-feedback condition. (3) We fitted a sigmoid from the ranked sequence of N distances to the ranked sequence of N experimental probabilities. (4) The sigmoid was applied to each perceptual distance in each condition in the model to directly calculate the probability of a "different" response. The same method was used to calculate the response probabilities in the expand test, except that the sigmoid was fitted using the expand test data from the model and the experiment.

7. This equivalence is approximate, of course, since we are ignoring factors like β_{21} (the receptive field size of L2 units) and β_{input} (the size of each input bump).

8. Such up-regulation of inhibition may also be present in the SLP. Since the native-language categories would be stable and robust, the level of excitatory input (and, consequently, the within-SLP inhibition) would be expected to be fairly stable; in the asymptotic case, the inhibitory levels may be considered to be fixed. Therefore, the regulation of inhibitory activity may selectively modulate FLP learning without affecting the slower scale learning in the SLP.

9. Another difference concerns the skew in the perceptual representations. In Guenther et al.'s (2004) model, training decreases the overall level of AM activity, but does not systematically skew the representation as such. Upon our view, training systematically skews the representations of near-prototypical stimuli (cf. Figure 2). However, with current imaging techniques, it is difficult to resolve whether a neural representation is skewed in some way.

10. Zhang et al. (2005) suggest on the basis of MEG data that Japanese listeners have greater cortical activity when listening to nonnative (/r/ and /l/) sounds in comparison with native (/b/ and /w/) sounds. However, the activity level was measured by the number of equivalent current dipole clusters, so the "greater activity" may be due to increased variability rather than a larger locus of activity. Significantly, the latency of the maximum MEG response was slower for nonnative sounds. With these caveats, Zhang et al.'s results are generally consistent with our hypothesis that native-sound perception involves greater and more focused neural activity (this idea is also similar to Zhang et al.'s own proposal of "neural commitment").

(Manuscript received October 15, 2005;
revision accepted for publication April 26, 2006.)