Perception & Psychophysics
2000, 62 (4), 843-867

Neural network models of categorical perception

R. I. DAMPER and S. R. HARNAD
University of Southampton, Southampton, England

Studies of the categorical perception (CP) of sensory continua have a long and rich history in psychophysics. In 1977, Macmillan, Kaplan, and Creelman introduced the use of signal detection theory to CP studies. Anderson and colleagues simultaneously proposed the first neural model for CP, yet this line of research has been less well explored. In this paper, we assess the ability of neural-network models of CP to predict the psychophysical performance of real observers with speech sounds and artificial/novel stimuli. We show that a variety of neural mechanisms are capable of generating the characteristics of CP. Hence, CP may not be a special mode of perception but an emergent property of any sufficiently powerful general learning system.

Studies of the categorical perception (CP) of sensory continua have a long and rich history. For a comprehensive review up until a decade ago, see the volume edited by Harnad (1987). An important question concerns the precise definition of CP. According to the seminal contribution of Macmillan, Kaplan, and Creelman (1977), "the relation between an observer's ability to identify stimuli along a continuum and his ability to discriminate between pairs of stimuli drawn from that continuum is a classic problem in psychophysics" (p. 452). The extent to which discrimination is predictable from identification performance has long been considered the acid test of categorical—as opposed to continuous—perception.

Continuous perception is often characterized by (approximate) adherence to Weber's law, according to which the difference limen is a constant fraction of stimulus magnitude. Also, discrimination is much better than identification:[1] Observers can make only a small number of identifications along a single dimension, but they can make relative (e.g., pairwise) discriminations between a much larger number of stimuli (G. A. Miller, 1956). By contrast, CP was originally defined (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Hoffman, & Griffith, 1957) as occurring when the *grain* of discriminability coincided with (and was, hence, predictable from) identification grain (i.e., when subjects could only discriminate *between* identifiable categories, not *within* them). Consequently, in CP, the discrimination of equally separated stimulus pairs as a function of stimulus magnitude is nonmonotonic, peaking at a category boundary (Wood, 1976) defined as a 50% point on a sharply rising or falling psychometric labeling function.

In fact, there are *degrees* of CP. The strongest case, in which identification grain would be equal to discrimination grain, would mean that observers could only discriminate what they could identify: One just-noticeable difference would be the size of the whole category. This (strongest) criterion has *never* been met empirically: It has always been possible to discriminate within the category, and not just between. In his 1984 review, Repp calls the coincidence of the point of maximum ambiguity in the identification function with the peak in discrimination "the essential defining characteristic of categorical perception" (p. 253). Macmillan et al. (1977) generalize this, defining CP as "the equivalence of discrimination and identification measures," which can take place "in the absence of a boundary effect since there need be no local maximum in discriminability" (p. 453).

In the same issue of *Psychological Review* in which the influential paper of Macmillan et al. (1977) appeared, Anderson, Silverstein, Ritz, and Jones (1977) applied to CP a trainable neural network model for associative memory (the brain-state-in-a-box [BSB] model), based jointly on neurophysiological considerations and linear systems theory. Simulated labeling (identification) and ABX discrimination curves were shown to be very much like those published in the experimental psychology literature. In the present paper, we will refer to such network models as exhibiting *synthetic* CP. Unlike psychophysical studies of real (human and animal) subjects, which have maintained an active profile, synthetic CP (with some important exceptions, reviewed below) has remained largely unexplored,[2] despite the vast and continuing upsurge of interest in neural models of perception and cognition be-

We are grateful to David Wykes, who programmed the brain-state-in-a-box neural simulation. Neil Macmillan kindly clarified for us some technical points relating to the psychophysics of the ABX task. The perceptive, critical comments of Robert Goldstone, Stephen Grossberg, Keith Kluender, Neil Macmillan, Dom Massaro, Dick Pastore, David Pisoni, Philip Quinlan, Mark Steyvers, Michael Studdert-Kennedy, and Michel Treisman on an earlier draft of this paper enabled us to improve it markedly. The VOT stimuli used here were produced at Haskins Laboratories, New Haven, Connecticut, with assistance from NICHD Contract NO1-HD-5-2910. Thanks to Doug Whalen for his time and patience. Correspondence concerning this article should be addressed to R. I. Damper, Image, Speech and Intelligent Systems (ISIS) Research Group, Bldg. 1, Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, England (e-mail: rid@ecs.soton.ac.uk).

ginning 10 or so years ago, as documented in the landmark volume of Rumelhart and McClelland (1986) and updated by Arbib (1995). An advantage of such computational models is that, unlike real subjects, they can be "systematically manipulated" (Wood, 1978, p. 583) to uncover their operational principles, a point made more recently by Hanson and Burr (1990), who write that "connectionist models can be used to explore systematically the complex interaction between learning and representation" (p. 471).

The purpose of the present paper is to assess neural net models of CP, with particular reference to their ability to simulate the behavior of real observers. As with any psychophysical model, the points in which synthetic CP agrees with observation show that real perceptual and cognitive systems could operate on the same principles as those embodied in the model. Where real and synthetic behaviors differ, this can suggest avenues for new experimentation.

The remainder of this paper is structured as follows. In the next section, we outline the historical development of theories of CP and its psychophysical basis. We then review various neural net models for synthetic CP. These have mostly considered artificial or novel continua, whereas experimental work with human subjects has usually considered speech (or more accurately, synthetic, *near-speech* stimuli), especially syllable-initial stop consonants. Thus, we describe the use of two rather different neural systems to model the perception of stops. The application of signal detection theory (SDT) to synthetic CP is then considered. Finally, the implications of the results of connectionist modeling of CP are discussed, before we present our conclusions and identify future work.

CHARACTERIZATION OF CATEGORICAL PERCEPTION

CP is usually defined relative to some theoretical position. Views of CP have accordingly evolved in step with perceptual theories. However CP is defined, the relation between discrimination and identification remains a central one. At the outset, we distinguish categorical *perception* from *mere* categorization (*sorting*) in that there is no warping of discriminability, rated similarity, or interstimulus representation distance (i.e., compression within categories and separation between) in the latter. Also, CP can be *innate*, as in the case of color vision (see, e.g., Bornstein, 1987), or *learned* (see, e.g., Goldstone, 1994, 1998).

Early Characterizations From Speech Categorical Perception

The phenomenon of CP was first observed and characterized in the seminal studies of the perception of synthetic speech at Haskins Laboratories, initiated by Liberman et al. (1957; see Liberman, 1996, for a comprehensive historical review). The impact of these studies on the field has been tremendous. Massaro (1987b) writes, "the study of speech perception has been almost synonymous with the study of categorical perception" (p. 90).

Liberman et al. (1957) investigated the perception of syllable-initial stop consonants (/b/, /d/, and /g/) varying in place of articulation, cued by second formant transition. Liberman, Delattre, and Cooper (1958) went on to study the voiced/voiceless contrast cued by first formant (*F*1) cutback, or voice onset time (VOT).[3] In both cases, perception was found to be categorical, in that a steep labeling function and a peaked discrimination function (in an ABX task) were observed, with the peak at the phoneme boundary corresponding to the 50% point of the labeling curve. Fry, Abramson, Eimas, and Liberman (1962) found the perception of long, steady-state vowels to be much "less categorical" than stop consonants.

An important finding of Liberman et al. (1958) was that the voicing boundary depended systematically on place of articulation. In subsequent work, Lisker and Abramson (1970) found that, as the place of articulation moves back in the vocal tract from bilabial (for a /ba–pa/ VOT continuum) through alveolar (/da–ta/) to velar (/ga–ka/), so the boundary moves from about 25-msec VOT through about 35 msec to approximately 42 msec. Why this should happen is uncertain. For instance, Kuhl (1987) writes that "we simply do not know why the boundary 'moves'" (p. 365). One important implication, however, is that CP is more than merely bisecting a continuum; otherwise, the boundary would be at midrange in all three cases.

At the time of the early Haskins work and for some years thereafter, CP was thought to reflect a mode of perception special to speech (e.g., Liberman et al., 1967; Liberman et al., 1957; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970), in which the listener somehow made reference to production. It was supposed that an early and irreversible conversion of the continuous sensory representation into a discrete, symbolic code subserving both perception and production (*motor theory*) took place. Thus, perception of consonants is supposedly *more categorical* than that of steady-state vowels because the articulatory gestures that produce the former are more discrete than the relatively continuous gestures producing the latter.

Although there is little or no explicit mention of Weber's law in early discussions of motor theory, its violation is one of the aspects in which CP was implicitly supposed to be special. Also, at that time, CP had not been observed for stimuli other than speech, a situation that was soon to change. According to Macmillan, Braida, and Goldberg (1987), however, "all . . . discrimination data require psychoacoustic explanation, whether they resemble Weber's Law, display a peak, or are monotonic" (p. 32). Despite attempts to modify it suitably (e.g., Liberman, 1996; Liberman and Mattingly, 1985, 1989), the hypothesis that CP is special to speech has been unable to bear the weight of accumulating contrary evidence.

One strong line of contrary evidence comes from psychoacoustic investigations, notably that of Kuhl and

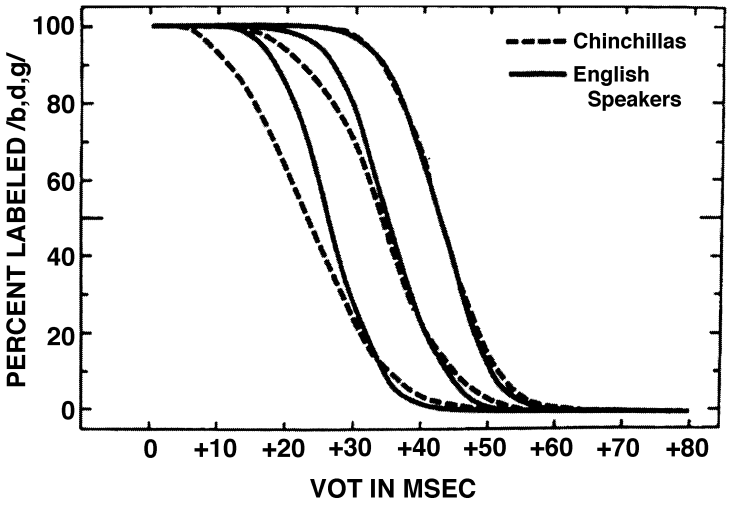

Figure 1. Mean identification functions obtained for bilabial, alveolar, and velar synthetic voice onset time (VOT) series for human listeners and chinchillas. Smooth curves have been fitted to the raw data points by probit analysis. From "Speech Perception by the Chinchilla: Identification Functions for Synthetic VOT Stimuli," by P. K. Kuhl and J. D. Miller, 1978, *Journal of the Acoustical Society of America*, *63*, p. 913. Copyright 1978 by Acoustical Society of America. Reprinted with permission.

Miller (1978), using nonhuman animal listeners who "by definition, [have] no phonetic resources" (p. 906). These workers trained 4 chinchillas to respond differentially to the 0- and 80-msec endpoints of a synthetic VOT continuum, as developed by Abramson and Lisker (1970). They then tested their animals on stimuli drawn from the full 0–80 msec range. Four human listeners also labeled the stimuli for comparison. Kuhl and Miller found "no significant differences between species on the absolute values of the phonetic boundaries . . . obtained, but chinchillas produced identification functions that were slightly, but significantly, less steep" (p. 905). Figure 1 shows the mean identification functions obtained for bilabial, alveolar, and velar synthetic VOT series (Kuhl & Miller's Figure 10). In this figure, smooth curves have been fitted to the raw data points (at 0, 10, 20, . . . 80 msec). Subsequently, working with macaques, Kuhl and Padden (1982, 1983) confirmed that these animals showed increased discriminability at the phoneme boundaries. Although animal experiments of this sort are methodologically challenging and there have been difficulties in replication (e.g., Howell, Rosen, Laing, & Sackin, 1992, working with chinchillas), the convergence of human and animal data in this study has generally been taken as support for the notion that general auditory processing and/or learning principles underlie this version of CP.

The emerging classical characterization of CP has been neatly summarized by Treisman, Faulkner, Naish, and Rosner (1995) as encompassing four features: "a sharp category boundary, a corresponding discrimination peak, the predictability of discrimination function from identification, and resistance to contextual effects" (p. 335). These authors go on to critically assess this characterization, referring to "the unresolved difficulty that identification data usually predict a lower level of discrimination than is actually found" (pp. 336–337) as, for example, in the work of Liberman et al. (1957), Macmillan et al. (1977), Pastore (1987b), and Studdert-Kennedy et al. (1970). They also remark on the necessity of qualifying "Studdert-Kennedy et al.'s claim that context effects [and other sequential dependencies] are weak or absent in categorical perception" (p. 337; see also Healy & Repp, 1982). We will take the classical characterization of CP to encompass only the first three aspects identified above, given the now rather extensive evidence for context effects and sequential dependencies (e.g., Brady & Darwin, 1978; Diehl, Elman, & McCusker, 1978; Diehl & Kluender, 1987; Repp & Liberman, 1987; Rosen, 1979) that can shift the category boundary.[4]

Signal Detection and Criterion-Setting Theories

The pioneering work at Haskins on phonetic categorization took place at a time when psychophysics was dominated by threshold models and before the influence of SDT (Green & Swets, 1966; Macmillan & Creelman, 1991) was fully felt. Analyses based on SDT differ from classical views of CP in two respects. First, SDT clearly separates measures of sensitivity (d') from measures of response bias (β). Second, the two views differ in the details of how discrimination is predicted from identification, with labeling playing a central role in both aspects of performance in the classical view. We deal with the latter point in some detail below; brief remarks on the first point follow immediately.

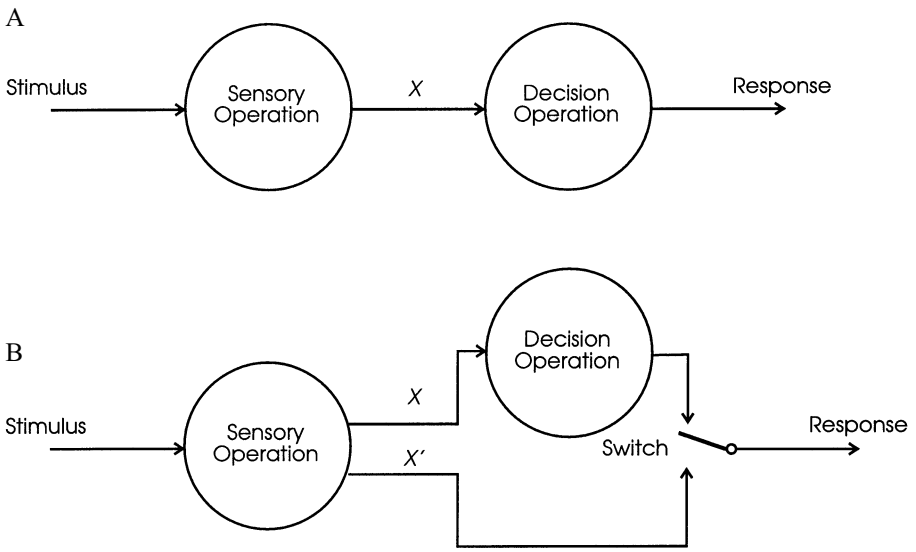

Figure 2. (A) The transformation of stimulus to response can be seen as a two-stage process of a sensory operation followed by a decision operation. This is consistent with signal detection theory's separation of sensitivity and response bias measures. Redrawn from Massaro, 1987a. (B) In Fujisaki and Kawashima's (1969, 1970, 1971) dual-process theory, there are two routes from sensory processing to decision: one continuous (X) and the other discrete (X').

Figure 2A (after Massaro, 1987a) shows the transformation of stimulus to response in an identification or a discrimination task as a two-stage process: a sensory operation followed by a decision operation. This is obviously consistent with SDT's separation of sensitivity factors (having to do with sensory perception) and response bias (having to do with decision processes). The importance of this in the present context is that classical notions of CP are ambiguous about which of the representations are categorical: The information passed between the sensory and the decision processes (labeled X in the figure) could be categorical or continuous. In the latter case, this still allows the response to be categorical, but this is not CP on Massaro's view, because the categorization does not occur at the sensory/perceptual stage: He prefers the term *categorical partition*. Emphasizing this distinction, Massaro (1987b) writes, "I cannot understand why categorization behavior was (and continues to be) interpreted as evidence for categorical perception. It is only natural that continuous perception should lead to sharp category boundaries along a stimulus continuum" (p. 115). (See also Hary & Massaro, 1982, and the reply of Pastore, Szczesiul, Wielgus, Nowikas, and Logan, 1984.)

According to SDT, X represents a *continuous* decision variate. In the simplest case, there are two kinds of presentation (historically called *signal* and *signal-plus-noise*), and X is unidimensional. The two classes of presentation give rise to two normal distributions of equal variance, one with a mean of zero and the other with a mean of d'. Stimuli are then judged to be from one class or the other, according to whether they give rise to an X value that is greater than or less than some internal criterion. However, as was detailed by Macmillan et al. (1977) and reviewed below, the paradigms used in the study of CP have generally been more complex than this simple case.

Is the internal criterion fixed, or can it shift as experience changes? This question has been addressed in the recent work of Treisman et al. (1995), who applied the earlier criterion-setting theory (CST) of Treisman and Williams (1984) to CP. According to CST, a sensory system resets the response criterion between each trial according to "the latest information available to it, about its own sensory performance and the environment" (p. 337), leading to sequential dependencies. The relevance to CP had been noted by Elman (1979), who suggested that consonant adaptation effects might be due to such criterion shifts. When applied to ABX discrimination, CST "is shown to fit the data from the literature" (Treisman et al., 1995, p. 334), in that a peak occurs at the category boundary. This is essentially because CST shares the basic assumptions of the classical Haskins model (p. 345), which also predicts (from labeling) a peak, as will be described below. Moreover, the absolute value of the observed discrimination performance is close to that predicted by CST. This is not the case with the Haskins model, which predicts a lower performance than is actually observed, as will be discussed immediately below. The better fit achieved by CST, relative to the Haskins model, is attributed to the former's additional criterion-setting assumptions.

Prediction of Discrimination From Identification

In the classical Haskins view, discrimination in an ABX task (as traditionally used in CP studies) is based on *covert* labeling. First, A is labeled covertly (in the sense

that the subject is not required to report this judgment to the investigator, as in overt identification), then B, then X: If the A and B labels are different, the subject responds *X is A* or *X is B* according to X's label; otherwise, the subject guesses. On this basis, ABX discrimination is predictable from identification. Indeed, one of the criticisms of this paradigm (e.g., Massaro & Oden, 1980; Pisoni & Lazarus, 1974) is that it *promotes* identification/labeling behavior, thereby arguably promoting categorization behavior also. For judgments involving just two categories, where the prior probability of each is equal, the proportion correct in discrimination is predicted as

$$P(C) = .5[1 + (p_A - p_B)^2], \quad (1)$$

where p_A is the probability of identifying the A stimulus as one of the two categories, p_B is the probability of identifying the B stimulus as that same category, and the guessing probability is .5 (Liberman et al., 1957; Macmillan et al., 1977). It is well known that this model predicts discrimination that is almost invariably lower than that observed. CP theorists have usually played down this discrepancy by emphasizing the correlation between the predicted and the observed curves—that is, their similar, nonmonotonic shape and the fact that they peak at approximately the same (boundary) point.

Massaro (1987b) writes, "for some reason, the discrepancy has never been a deterrent for advocates of categorical perception nor a central result for any alternative view" (p. 91). However, the dual-process model of Fujisaki and Kawashima (1969, 1970, 1971) does indeed effectively take this discrepancy as the basis of an alternative view, in which both a continuous (*auditory*) and a categorical (*phonetic*) mode of processing coexist (Figure 2B). If the subject fails to label A and B differently via the categorical route, then, rather than guessing, the continuous (but decaying) representations of A and B are consulted. According to Macmillan et al. (1977, p. 454), the extent to which Equation 1 underestimates discrimination determines the weight to be given to each process so as to fit the data best. They criticize dual-process theory for "its embarrassing lack of parsimony" (p. 467), however, in that everything that can be done via the discrete route (and more) can also be achieved via the continuous route. The theory does, however, have other strengths. It can explain, for instance, the effect that memory requirements of the experimental procedure have on CP on the basis that the two processes have different memory decay properties.

Macmillan et al. (1977) point out that the Haskins model is tacitly based on low-threshold assumptions,[5] arguing that mere correlation between observed discrimination and that predicted from identification is inadequate support for the notion of CP. By contrast, they characterize CP, on the basis of SDT, in terms of the equivalence of discrimination and identification. The essential defining characteristic of CP is then considered to be the equivalence of identification d', found by using the approach proposed by Braida and Durlach (1972) for auditory intensity perception, and discrimination d'. The Braida and Durlach model assumes a distribution corresponding to each point on the continuum and then finds a d' for each adjacent pair of distributions. If we can find a d' corresponding to the same pair of distributions in ABX discrimination, these two sensitivity measures should be equal, if discrimination is indeed predictable from identification.

To avoid the low-threshold assumptions of a discrete set of internal states, Macmillan et al. (1977) extended Green and Swets' earlier (1966) derivation of d' from yes–no and two-interval forced-choice (2IFC) psychophysical tasks to the somewhat more complicated ABX task. It was analyzed (pp. 458–459) as a 2IFC subtask (to determine whether the standards are in the order ⟨AB⟩ or ⟨BA⟩), followed by a yes–no subtask.[6] This is described as "a continuous (SDT) model for categorical perception" (p. 462). This view of the importance of continuous information to CP is gaining ground over the classical characterization of CP. For instance, Treisman et al. (1995) state that "CP resembles standard psychophysical judgments" (p. 334), whereas Takagi (1995) writes, "in fact, the signal detection model is compatible with both categorical and continuous patterns of identification/discrimination data" (p. 569).

NEURAL MODELS OF CATEGORICAL PERCEPTION: A REVIEW

In this section, we present a historical review of synthetic CP.

The Brain-State-in-a-Box

Early neural models of categorical perception were essentially based on associative memory networks—one of the few kinds of net attracting any kind of interest in the "dark ages" (see note 2) before the discovery of the error back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986). (See Hinton & Anderson, 1981, and Kohonen, 1977, for extensive contemporary reviews of parallel models of associative memory, and Anderson, 1995, for a more recent introductory treatment.) This is quite a natural model for CP in many ways. An associative net is addressed with some partial or noisy pattern and retrieves the corresponding noise-free canonical pattern, or prototype. This is akin to a pure or classical form of CP whereby a nonprototypical stimulus is replaced in memory by its prototype (from which it is consequently indistinguishable).

We will take Anderson, Silverstein, et al.'s (1977) paper as the starting point for our review of neural models of CP. We note, however, that this selection may be contentious. Grossberg (1968a, 1968b, 1969), for example, also has a legitimate claim to be the originator of this line of research with his very early papers on neural models of psychological functions, although Anderson's work on associative memory dates back at least as far (viz., Anderson, 1968; see Carpenter, Grossberg, & Rosen, 1991a, 1991b, and Grossberg, 1987, for more recent develop-

ments.) We prefer the Anderson, Silverstein, et al. (1977) model, because of its greater simplicity and perspicacity and its more direct and obvious usefulness in modeling human psychophysical data.

Anderson, Silverstein, et al. (1977) consider networks of neurons,[7] which "are simple analog integrators of their inputs" (p. 416). They extend the earlier work mentioned above (e.g., Anderson, 1968) in two main ways. It had previously been assumed (p. 413) that (1) nervous system activity could be represented by the pattern of activation across a group of cells, (2) different memory traces make use of the same synapses, and (3) synapses associate two patterns by incrementing synaptic weights in proportion to the product of pre- and postsynaptic activities.

The form of learning implied in the third assumption is, in effect, correlational and has been called *Hebbian* by many workers. Since the neurons have linear activation functions, a form of linear correlation is computed, making the net amenable to analysis with linear systems theory as follows.

Suppose N-dimensional input pattern vectors $\mathbf{f}_i$ are to be associated with M-dimensional output pattern vectors $\mathbf{g}_i$. A net is created with N input units and M output units. In accordance with the second assumption above, $\mathbf{f}$ and $\mathbf{g}$ are to be represented by the patterns of activation across the input and output units, respectively. Then, according to the learning scheme, the $(M \times N)$ connection matrix $\mathbf{A}$ of synaptic weights between the two sets of units is incremented by

$$\mathbf{A}_i = \mathbf{g}_i \mathbf{g}_i^T, \tag{2}$$

where T denotes the vector transpose. In this way, the overall connectivity matrix is determined as $\mathbf{A} = \sum_i \mathbf{A}_i$, summed over all I input patterns. If all inputs are mutually orthogonal, the output for any $\mathbf{f}_k$ will be

$$\begin{aligned}\mathbf{A}\mathbf{f}_k &= \sum_{i=1}^{I} \mathbf{A}_i \mathbf{f}_k \\ &= \mathbf{A}_k \mathbf{f}_k + \sum_{i \neq k} \mathbf{A}_i \mathbf{f}_k \\ &= \mathbf{g}_k \mathbf{f}_k^T \mathbf{f}_k + \sum_{i \neq k} \mathbf{g}_i \mathbf{f}_i^T \mathbf{f}_k \text{ by Equation 2} \\ &\propto \mathbf{g}_k ,\end{aligned}$$

since, by the definition of orthogonality,

$$\mathbf{f}_i^T \mathbf{f}_j = \begin{cases} \|\mathbf{f}_i\|^2 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the system operates as a perfect associator in this case: The direction of the output vector is identical to that of the associated input vector. (The length, however, is modified by the length of the input vector and will also depend on the number of repetitions of that input in accordance with Equation 2.) When the inputs are not orthogonal, the net will produce noise, as well as the correct output, but it will still be "quite usable" (Anderson, Silverstein, et al., 1977, p. 417).

To convert this linear pattern-association net into a model of CP, Anderson, Silverstein, et al. (1977) made two extensions. The first was to discard the M distinct output units and to introduce positive feedback from the set of N input neurons onto itself. The $(N \times N)$ matrix $\mathbf{A}$ (which they now call the *feedback matrix*) is made symmetric in this case, so that the synaptic weight between units i and j is equal to that between units j and i: $a_{ij} = a_{ji}$. For the case of arbitrary (nonorthogonal) inputs, it is shown (Anderson, Silverstein, et al., 1977, p. 424) that (provided their average is zero) the inputs are a linear combination of the eigenvectors of the feedback matrix $\mathbf{A}$ and that all eigenvalues are positive.

The introduction of positive feedback makes the system potentially unbounded, in that activations can now grow without limit. The second extension overcomes this problem by allowing the individual neurons to saturate at an activation of $\pm C$. That is, the activation function of each neuron is linear-with-saturation. Thus, in use, all the units are eventually driven into saturation (either positive or negative in sense), and the net has stable states corresponding to some (possibly all) of the corners of a hypercube (*box*) in its N-dimensional state space. (Of course, not all corners are necessarily stable.) For this reason, the model was called *brain-state-in-a-box*. Considered as vectors in the state space, these corners are the eigenvectors of $\mathbf{A}$ and can be identified, in psychological terms, with the *distinctive features* of the system (Anderson, Silverstein, et al., 1977, p. 425). For each such stable state, there is a region of attraction in state space such that, if an input initially produces an output in this region, that output will evolve over time to reach that stable state, where it will remain.

Used as a model of CP, the inputs (i.e., the $\mathbf{f}_i$) are associated with themselves during training—that is, during computation of the $(N \times N)$ matrix $\mathbf{A}$. This is an essentially unsupervised operation. However, if the training patterns are labeled with their pattern class, the corners of the box can be similarly labeled according to the input patterns that they attract. (There is, of course, no guarantee that all the corners will be so labeled. Corners that remain unlabeled correspond to *rubbish states*, in the jargon of associative networks.) Thereafter, an initial noisy input (consisting of a linear sum of eigenvectors) within the region of attraction of a labeled corner will evoke an output that is a canonical or prototypical form corresponding to the eigenvector of the input with the largest eigenvalue. Anderson, Silverstein, et al. (1977, pp. 430–433) present a simulation of CP in which their neural model performed two-class identification and ABX discrimination tasks. The two prototypes (eigenvectors) were the eight-dimensional orthogonal vectors of length two in the directions (1,1,1,1, −1,−1,−1,−1) and (1,1,−1,−1,1,1,−1,−1), respectively (Figure 3A). These were used to set the weights as detailed above. Inputs to the model then consisted of 100 repetitions of 16 length-one vectors equally spaced between the prototype eigenvectors, with added zero-mean Gaussian noise according to one of four conditions: The standard deviation (*SD*) of the noise was 0.0, 0.1, 0.2, or 0.4.

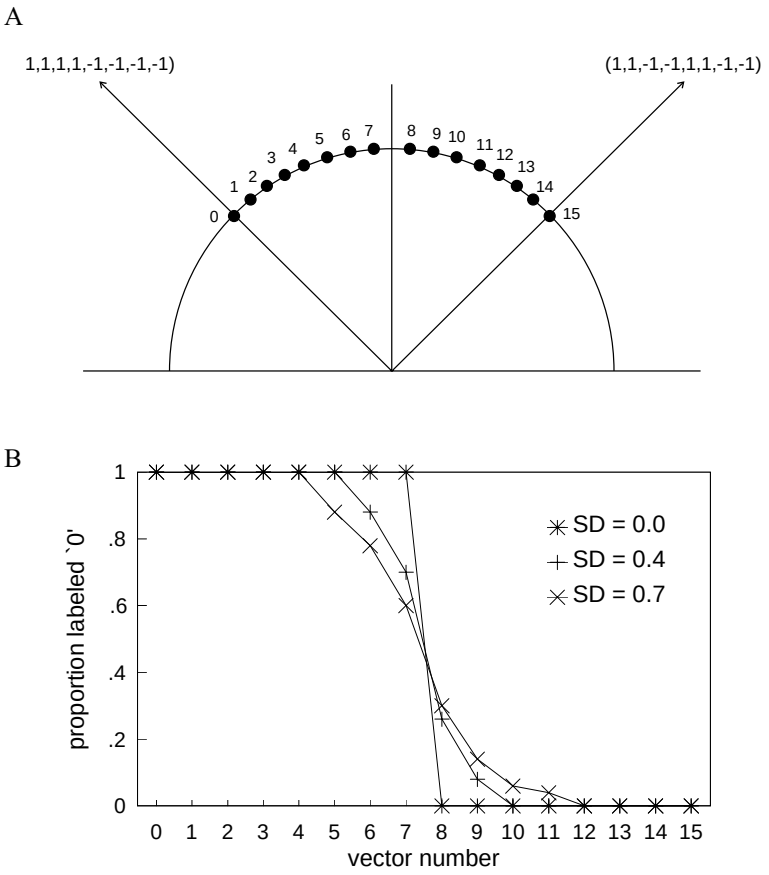

Figure 3. (A) An eight-dimensional system, with two orthogonal length-two eigenvectors, used in the simulation of categorical perception on the basis of the brain-state-in-a-box neural model. Inputs consisted of 16 equally spaced length-one vectors, as shown, with added noise. Redrawn from Anderson, Silverstein, Ritz, and Jones (1977). (B) Response of model for the simulated identification task with 1,000 presentations at each noise condition: The standard deviation was 0.0, 0.4, or 0.7.

We have replicated Anderson, Silverstein, et al.'s (1977) simulation. Weights were calculated from the inner product of each of the two training patterns with itself, added to produce the feedback matrix in accordance with Equation 2. Testing used 1,000 presentations under three different noise conditions. During testing, the activation of each neuron was computed as

$$act_i(t) = \alpha(extinput_i(t)) + \beta(intinput_i(t)), \qquad (3)$$

where $extinput_i$ and $intinput_i$ are, as their names clearly suggest, the external input to unit i and the internal (feedback) input to the same unit. A decay term,

$$\Delta act_i(t) = \alpha(extinput_i(t)) + \beta(intinput_i(t)) - (decay)act_i(t),$$

can be incorporated into the model, which tends to restore activation to a resting level of zero. Throughout this work, *decay* was set to 1 so that the activation is given simply by Equation 3.

For the replication of Anderson, Silverstein, et al.'s (1977) simulation, the external scale factor α and the internal scale factor β were both set at .1. The saturation limit for the neurons was set at $C = \pm 1$. Self-connections between neurons were allowed. We also found it necessary to use rather more noise power than Anderson, Silverstein, et al. did. We believe this is because our use of 1,000 test patterns (in place of Anderson, Silverstein, et al.'s 100) makes our results less affected by small-sample effects. Thus, our noise conditions were $SD = 0.0$, 0.4, and 0.7.

In all noise-free cases, the system converged to one of its two stable, saturating states for all 1,000 inputs. For the added noise conditions, there was only a very small likelihood of convergence to an unlabeled corner (rubbish state). This occurred for approximately 1% of the inputs when $SD = 0.4$ and for about 6% when $SD = 0.7$. Fig-

ure 3B shows the identification results obtained by noting the proportion of inputs that converged to the saturating state corresponding to endpoint 0. For the no-noise condition, categorization was perfect with the class boundary at the midpoint between the prototypes. For the noise conditions, $SD = 0.4$ and $SD = 0.7$, the labeling curves were very reasonable approximations to those seen in the classical CP literature. Overall, this replicated the essential findings of Anderson, Silverstein, et al. (1977).

Consider next the ABX discrimination task. Anderson, Silverstein, et al. (1977) considered two inputs to the net to be discriminable if they converged to different stable states. (Note that since Anderson, Silverstein, et al. were considering a simple two-class problem with convergence to one or the other of the two labeled states and no rubbish states, they were never in the situation of having A, B, and X all covertly labeled differently, as can conceivably happen in reality.) If they converged to the same stable state, a guess was made with a probability of .5, in accordance with Equation 1. This means that discrimination by the net was effectively a direct implementation of the Haskins model. Indeed, Anderson, Silverstein, et al. observed a distinct peak at midrange for their intermediate-noise condition, just as in classical CP. Finally, they obtained simulated *reaction times* by noting the number of iterations required to converge to a stable, saturating state. As in classical CP (e.g., Pisoni & Tash, 1974), there was an increase in reaction time for inputs close to the category boundary for the intermediate-noise condition, relative to inputs more distant from the boundary. Again, we have replicated these findings (results not shown).

In support of the assertion that the model is "quite usable" when the inputs are not orthogonal, Anderson (1977, pp. 78–83) presents an example in which the BSB model was used to categorize vowel data (see also Anderson, Silverstein, & Ritz, 1977). Twelve Dutch vowels were represented by eight-dimensional vectors, each element measuring the energy within a certain frequency band of an average, steady-state vowel. It is highly unlikely that these inputs were mutually orthogonal; yet, "when learning ceased, each vowel was assigned to a different corner" (p. 81). Indeed, as was mentioned earlier, nonorthogonality can act as noise, thus preventing (unrealistic) perfect categorization.

Anderson, Silverstein, et al. (1977) conjecture that positive feedback, saturation, and synaptic learning were "responsible for the interesting [categorization] effects in our simulations" (p. 433). With the benefit of hindsight, however, we now know (on the basis of the extensive review material and the new results below) that synthetic categorization can be obtained in a variety of neural models, even those lacking positive feedback and saturation. In this regard, the comments of Grossberg (1986) concerning saturation in the BSB model are apposite. He charged Anderson, Silverstein, et al. with introducing a homunculus as a result of their "desire to preserve the framework of linear systems theory." He continues: "No physical process is defined to justify the discontinuous change in the slope of each variable when it reaches an extreme of activity . . . The model thus invokes a homunculus to explain . . . categorical perception" (pp. 192–194).

In our view, however, a homunculus is an unjustified, implicit mechanism that is, in the worst case, comparable in sophistication and complexity to the phenomenon to be explained. By contrast, Anderson, Silverstein, et al. (1977) postulate an explicit mechanism (firing-rate saturation) that is both simple and plausible, in that something like it is a ubiquitous feature of neural systems. In the words of Lloyd (1989), "homunculi are tolerable provided they can ultimately be discharged by analysis into progressively simpler subunculi, until finally each micrunculus is so stupid that we can readily see how a mere bit of biological mechanism could take over its duties" (p. 205). Anderson, Silverstein, et al. go so far as to tell us what this "mere bit of biological mechanism" is—namely, rate saturation in neurons. (See Grossberg, 1978, and the reply thereto of Anderson & Silverstein, 1978, for additional discussion of the status of the nonlinearity in the Anderson, Silverstein, et al. BSB model; see also Bégin & Proulx, 1996, for a more recent commentary.) To be sure, the discontinuity of the linear-with-saturation activation function is biologically and mathematically unsatisfactory, but essentially similar behavior is observed in neural models with activation functions having a more gradual transition into saturation (as will be detailed below).

The TRACE Model

In 1986, McClelland and Elman produced a detailed connectionist model of speech perception that featured localist representations and extensive top-down processing, in addition to the more usual bottom-up flow of information. This model, TRACE, is now rather well known, so it will be described only briefly here. There are three levels to the full model, corresponding to the (localist) feature, phoneme, and word units. Units at different levels that are mutually consistent with a given interpretation of the input have excitatory connections, whereas those within a level that are contradictory have inhibitory connections—that is, processing is *competitive*.

Strictly speaking, TRACE is as much a model of lexical accessing as of speech perception per se, as the existence of the word units makes plain. McClelland and Elman (1986) assumed an input in terms of something like *distinctive features*, which sidesteps important perceptual questions about how the distinctive features are derived from the speech signal and, indeed, about whether this is an appropriate representation or not. In their 1986 *Cognitive Psychology* paper, McClelland and Elman describe TRACE II, which, they say, is "designed to account primarily for lexical influences in phoneme perception" using "mock speech" as input (p. 13). However, they also refer to TRACE I, which is "designed to address some of the challenges posed by the task of recognizing phonemes in real speech" and cite Elman and McClelland (1986) for further details. Unfortunately, TRACE I does not feature real speech input either.

Top-down effects are manifest through the lexical status (or otherwise) of words affecting (synthetic) phoneme perception and, thereby, (synthetic) feature perception also. Although TRACE has been used to simulate a variety of effects in speech perception, we concentrate here on its use in the modeling of CP.

An 11-step /g/–/k/ continuum was formed by interpolating the feature values—namely, VOT and the onset frequency of the first formant, *F*1. The endpoints of the continuum (Stimuli 1 and 11) were more extreme than prototypical /g/ and /k/, which occurred at Points 3 and 9, respectively. The word units were removed, thus excluding any top-down lexical influence, and all phoneme units other than /g/ and /k/ were also removed. Figure 4A shows the initial activations (at time step $t = 1$) at these two units as a function of stimulus number. As can be seen, there was a clear trend for the excitation (which was initially entirely bottom-up) to favor /g/ at a low stimulus number but /k/ at a high stimulus number. The two curves cross at Stimulus 6, indicating that this condition was maximally ambiguous (i.e., this was the phoneme boundary). However, the variation was essentially continuous rather than categorical, as is shown by the relative shallowness of the curves. By contrast, after 60 time steps, the two representations were as shown in Figure 4B. As a result of the mutual inhibition between the /g/ and /k/ units and, possibly, of the top-down influence of phoneme units on featural units also, a much steeper (more categorical) response was seen.

This appears to be a natural consequence of the competition between excitatory and inhibitory processing. Many researchers have commented on this ubiquitous finding. For instance, Grossberg (1986) states, "Categorical perception can . . . be anticipated whenever adaptive filtering interacts with sharply competitive tuning, not just in speech recognition experiments" (p. 239).

McClelland and Elman (1986) go on to model overt labeling of the phonemes, basing identification on a variant of Luce's (1959) choice rule. The result is shown in Figure 4C, which also depicts the ABX discrimination function. The choice rule involves setting a constant k (actually equal to 5), which acts as a free parameter in a curve-fitting sense. Quinlan (1991) accordingly makes the following criticism of TRACE: "Indeed, k determined the shape of the identification functions. . . . A rather uncharitable conclusion . . . is that the model has been fixed up to demonstrate categorical perception. . . . Categorical perception does not follow from any of the a priori functional characteristics of the net" (p. 151). It is also apparent that the obtained ABX discrimination curve is not very convincing, having a rather low peak relative to those found in psychophysical experiments.

We consider finally the relation between discrimination and identification in TRACE. McClelland and Elman (1986) point out that discrimination in *real* CP is better than that predicted from identification and that TRACE also "produces this kind of approximate categorical perception" (p. 47). The mechanism by which this happens is an interaction of the bottom-up activations produced by the speech input with the top-down activations. According to the authors, the former decay with time, but not entirely to zero, whereas the latter produce a more canonical representation with time but do not completely overwrite the input with its prototype (and the time course of these interactions gives a way of predicting the increase in reaction time for stimuli close to the category boundary). The authors remark on the practical difficulty of distinguishing between this feedback explanation and a dual-process explanation.

Back-Propagation

As is well known, the field of neural computing received a major boost with the discovery of the error back-propagation algorithm (Rumelhart et al., 1986) for training feedforward nets with hidden units, so-called multilayer perceptrons (MLPs). It is, therefore, somewhat surprising that back-propagation learning has not figured more widely in studies of synthetic CP. We have used this algorithm as the basis for modeling the CP of both speech (dealt with in the next section) and artificial stimuli (dealt with here).

Many workers (Baldi & Hornik, 1989; Bourland & Kamp, 1988; Elman & Zipser, 1988; Hanson & Burr, 1990) have observed that feedforward auto-associative nets[8] with hidden units effectively perform a principal component analysis of their inputs. Harnad, Hanson, and Lubin (1991) exploited auto-association training to produce a *precategorization* discrimination function. This was then reexamined after categorization training, to see whether it had warped. That is, a CP effect was defined as a decrease in within-category interstimulus distances and/or an increase in between-category interstimulus distances relative to the baseline of auto-association alone. The stimuli studied were artificial—namely, different representations of the length of (virtual) lines—and the net's task was to categorize these as *short* or *long*.

A back-propagation net with 8 input units, a single hidden layer of 2–12 units, and 8 or 9 output units was used. The eight different input lines were represented in six different ways, to study the effect of the *iconicity* of the input coding (i.e., how analogue, nonarbitrary, or structure preserving it was in relation to what it represented).[9] After auto-association training (using 8 output units), the trained weights between hidden layer and output layer were reloaded, the input to hidden layer weights were set to small random values, and training recommenced. The net was given a double task: auto-association (again) and categorization. For the latter, the net had to label Lines 1–4 (for instance) as *short* and 5–8 as *long*. This required an additional output, making 9 in this case.

Strong CP effects (with warping of similarity space in the form of increased separation across the category boundary and compression within a category) were observed for all the input representations: The strongest effect was obtained with the least iconic, most arbitrary (place) code. The categorization task was very difficult

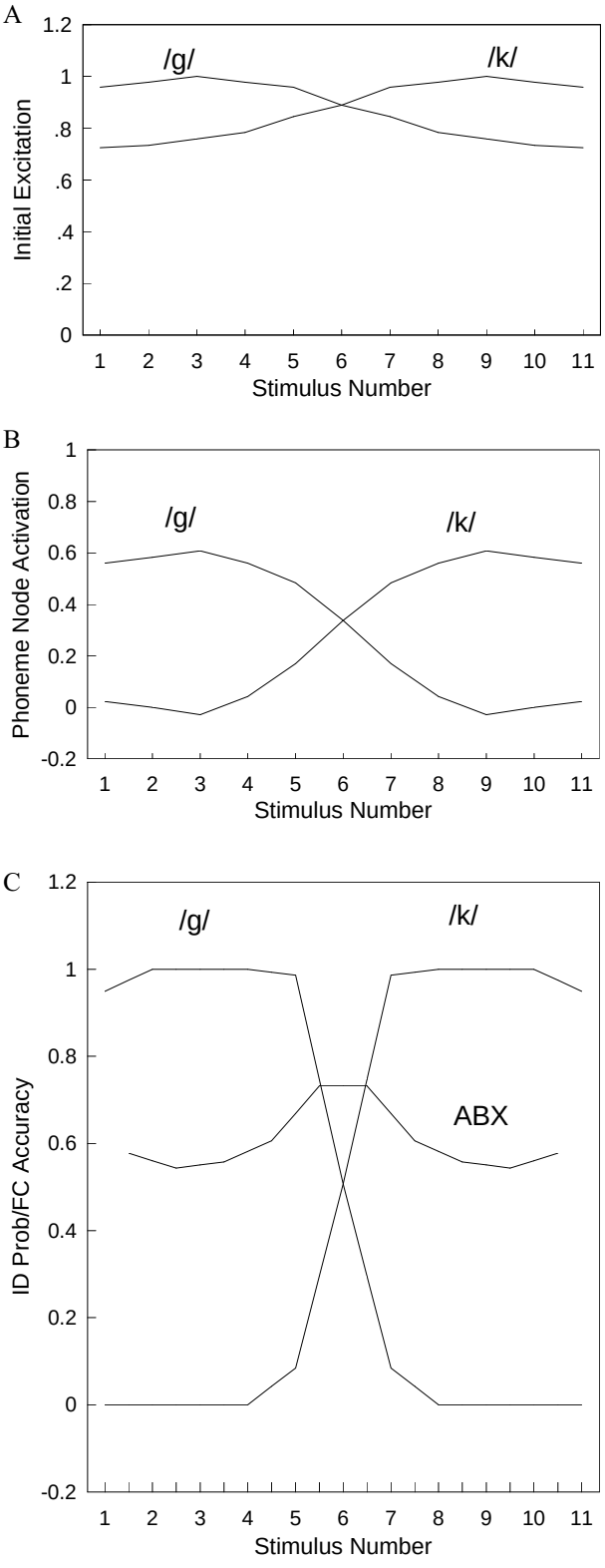

Figure 4. Categorical perception in the TRACE model. (A) Initial activation of the /g/ and /k/ units arising from bottom-up influence of the feature units, at time step $t = 1$. (B) Activations at time step $t = 60$. (C) Labeling functions after postprocessing, using Luce's choice model with $k = 5$, and ABX discrimination curve. Redrawn from McClelland and Elman (1986).

to learn with only two hidden units, $h = 2$. With more hidden units, however, the pattern of behavior did not change with increasing h (3–12). This is taken to indicate that CP was not merely a by-product of information compression by the hidden layer. Nor was CP a result of overlearning to extreme values, because the effect was present (albeit smaller) for larger values of the epsilon (ϵ) error criterion in the back-propagation algorithm. A test was also made to determine whether CP was an artifact of reusing the weights for the precategorization discrimination (auto-association) for the auto-association-plus-categorization nets. Performance was averaged over several precategorization nets and was compared with performance averaged over several different auto-association-plus-categorization nets. Again, although weaker and not always present, there was still evidence of synthetic CP.

A final test concerned iconicity and interpolation: Was the CP restricted to trained stimuli, or would it generalize to untrained ones? Nets were trained on auto-association in the usual way, and then, during categorization training, some of the lines were left untrained (say, Line 3 and Line 6) to see whether they would nevertheless warp in the "right" direction. Interpolation of the CP effects to untrained lines was found, but only for the coarse-coded representations.

A "provisional conclusion" of Harnad et al. (1991) was that "whatever was responsible for it, CP had to be something very basic to how these nets learned" (p. 70). In this and subsequent work (Harnad, Hanson, & Lubin, 1995), the time-course of the training was examined, and three important factors in generating synthetic CP were identified: (1) maximal interstimulus separation induced during auto-association learning, with the hidden-unit representations of each (initially random) stimulus moving as far apart from one another as possible; (2) stimulus movement to achieve linear separability during categorization learning, which undoes some of the separation achieved in the first factor above, in a way that promotes within-category compression and between-category separation; and (3) inverse-distance "repulsive force" at the category boundary, pushing the hidden-unit representation away from the boundary and resulting from the form of the (inverse exponential) error metric, which is minimized during learning.

One further factor—the iconicity of the input codings—was also found to modulate CP. The general rule is that the further the initial representation is from satisfying the partition implied in Points 1–3 above (i.e., the less iconic it is), the stronger the CP effect. Subsequently, Tijsseling and Harnad (1997) carried out a more detailed analysis, focusing particularly on the iconicity. Contrary to the report of Harnad et al. (1995), they found no overshoot, as in Point 2 above. They concluded, "CP effects usually occur with similarity-based categorization, but their magnitude and direction vary with the set of stimuli used, how [these are] carved up into categories, and the distance between those categories" (p. 268).

This work indicates that a feedforward net trained on back-propagation is able (despite obvious dissimilarities)

to replicate the essential features of classical CP much as the BSB model of Anderson, Silverstein, et al. (1977) did. There are, however, noteworthy differences. The most important is that Harnad et al.'s (1991) back-propagation nets were trained on intermediate (rather than solely on endpoint) stimuli. Thus, generalization testing is a more restricted form of interpolation. Also (because the feedforward net has no dynamic behavior resulting from feedback), reaction times cannot be quite so easily predicted as in Anderson, Silverstein, et al. (but see below).

Competitive Learning and Category-Detecting Neurons

Goldstone, Steyvers, and Larimer (1996) report on a laboratory experiment with human subjects in which stimuli from a novel dimension were categorically perceived. The stimuli were created by interpolating (*morphing*) seven curves between two randomly selected bezier endpoint curves. The dimension was novel in that the subjects were highly unlikely ever to have seen precisely those morphed shapes before. The major interest, in the context of this paper, is that Goldstone et al. also present a neural model (a form of radial-basis function net) that qualitatively replicates the behavioral results.

The model has a layer of hidden neurons that become specialized for particular stimulus regions, thereby acting as *category-detecting neurons* in the sense of Amari and Takeuchi (1978) or *feature-detecting neurons* in the sense of Schyns (1991). This is done by adjusting the input-to-hidden (or *position*) weights. Simultaneously, associations between hidden/detector neurons and output (*category*) units are learned by gradient descent. In addition to the feedforward connections from input-to-hidden and from hidden-to-output units, there is feedback from the category units, which causes the detector units to concentrate near the category boundary. This works by increasing the position-weight learning rate for detectors that are neighbors of a detector that produces an improper categorization. Note that the whole activity pattern of the hidden detectors determines the activity of the category nodes. This, in turn, determines the error and, thus, the learning rate. No single detector can determine the learning rate (Mark Steyvers, personal communication, July 9, 1997).

Goldstone et al. (1996) mention the similarity of the classification part of their model to ALCOVE (Kruschke, 1992). Like ALCOVE, the hidden nodes are radial-basis function units "activated according to the psychological similarity of the stimulus to the exemplar at the position of the hidden node" (p. 23). The essential difference is that Goldstone et al.'s exemplar nodes are topologically arranged and can move their position in input space through competitive learning of their position weights.

Simulations were performed with input patterns drawn from 28 points on the morphed continuum. (Two-dimensional gray-scale drawings of the curves were converted to Gabor filter representations describing the inputs in terms of spatially organized line segments.) There were 14 hidden exemplar/detector neurons and 2 output/category neurons. Like the experiments with the human subjects, the simulations involved learning two different classifications according to different cut-offs along the novel dimension. In one condition (*left split*), the cut-off (boundary) was placed between Stimuli 10 and 11; in the other condition (*right split*), it was placed between Stimuli 18 and 19. In both cases, classical CP was observed. Although Luce's (1959) choice rule is apparently used in the Goldstone et al. (1996) model, it seems that the k parameter, which was treated by McClelland and Elman (1986) as free in the TRACE model and was adjusted to give CP, is here treated as fixed (at unity). The labeling probability showed a characteristic warping, with its 50% point being at the relevant boundary. Discrimination between two stimuli was assessed by taking the Euclidean distance between their hidden-node activation patterns. This revealed a peak in sensitivity at or near the relevant category boundary.

Unfortunately, Goldstone et al. (1996) did not (and cannot) make a strict comparison of their human and simulation data, because of the different numbers of curves in the two continua studied. Recall that 7 morphed curves constituted the continuum for the experiments with human subjects, whereas a morphing sequence of 28 curves was used in the simulations. Such a comparison could have been very revealing for understanding synthetic CP. Nonetheless, there is sufficient coincidence of the *form* of their results in the two cases to show that neural nets can indeed make credible models of learned categorization.

The authors contrast their work with that of Anderson, Silverstein, et al. (1977) and Harnad et al. (1995). In these other approaches, they say, "each category has its own attractor,"[10] so that CP "occurs because inputs that are very close but fall into different categories will be driven to highly separated attractors" (Goldstone et al., 1996, p. 248). In their net, however, detectors congregate at the category boundary, and thus "small differences . . . will be reflected by [largely] different patterns of activity." These aspects of their work are presented as potentially advantageous. However, they seem to run counter to the prevailing view in speech CP research, according to which the paradigm "has overemphasized the importance of the phonetic boundary between categories" (Repp, 1984, p. 320) at the expense of exploring the internal structure of the categories in terms of anchors and/or prototypes (e.g., Guenter & Gjaja, 1996; Iverson & Kuhl, 1995; Kuhl, 1991; Macmillan, 1987; J. L. Miller, 1994; Volaitis & Miller, 1992; but see Lotto et al., 1998).

CATEGORIZATION OF STOP CONSONANTS BY NEURAL NETWORKS

From the foregoing review, it is apparent that (given the right encoding schema for the inputs) neural models of CP have no real problem replicating the classical observations of a sharp labeling function and a peaked discrimination function, at least for learned categorization.

Although there may sometimes be contrary suspicions (as when Luce's choice rule is used in the TRACE model or nets are trained to place the category boundary at a particular point on the input continuum), the effects are sufficiently robust across a variety of different architectures and approaches to support the claim that they reflect the emergent behavior of any reasonably powerful learning system (see below). With the exception of the vowel categorization work using the BSB model (Anderson, 1977; Anderson, Silverstein, et al., 1977), however, the neural models of synthetic CP reviewed thus far have all taken their inputs from artificial or novel dimensions, whereas the vast majority of real CP studies have used speech stimuli—most often, stop consonants (or, more correctly, simplified analogues of such sounds). Our goal in this section is, accordingly, to consider the categorization of stop consonants by a variety of neural models. As was mentioned earlier, an important aspect of the categorization of stop consonants is the shift of the category boundary with place of articulation. Thus, it is of considerable interest to ascertain whether neural models of CP reproduce this effect as emergent behavior.

Stimuli and Preprocessing

The stimuli used in this section were synthesized consonant–vowel syllables supplied by Haskins Laboratories and nominally identical to those used by Kuhl and Miller (1978), which were developed earlier by Abramson and Lisker (1970). Stimuli very much like these, if not identical to them, have been used extensively in studies of speech CP: they have become a *gold standard* for this kind of work. They consist of three series, digitally sampled at 10 kHz, in which VOT varies in 10-msec steps from 0 to 80 msec, simulating a series of English, prestressed, bilabial (/ba–pa/), alveolar (/da–ta/), and velar (/ga–ka/) syllables. Each stimulus began with a release burst, and the two acoustic variables of aspiration duration and *F*1 onset frequency were then varied simultaneously in order to simulate the acoustic consequences of variation in VOT. Strictly, then, the VOT continuum is not unidimensional. However, as was mentioned in note 3, these two variables have often been thought to be perfectly correlated.

The stimuli were preprocessed for presentation to the various nets, using a computational model of the peripheral auditory system (Pont & Damper, 1991). The use of such sophisticated preprocessing obviously requires some justification. We know from above that the iconicity of the input representation to the network is important: the closer the representation to that "seen" by the real observer the better. Also, there has long been a view in the speech research literature that CP reflects some kind of "restructuring of information" (Kuhl & Miller, 1978, p. 906) by the auditory system in the form of processing nonlinearities. We wished, accordingly, to find correlates of CP in the neural activity of the auditory system, following Sinex and McDonald (1988), who write: "It is of interest to know how the tokens from a VOT continuum are represented in the peripheral auditory system, and whether [they] tend to be grouped in a way which predicts the psychophysical results" (p. 1817). Also, as a step toward understanding the acoustic-auditory restructuring of information, we wished to discover the important acoustic features that distinguish initial stops. In the words of Nossair and Zahorian (1991), who used automatic speech recognition techniques for this purpose, "Such features might be more readily identifiable if the front-end spectral processing more closely approximated that performed by the human auditory system" (p. 2990). Full details of the preprocessing are described elsewhere (Damper, Pont, & Elenius, 1990). Only a brief and somewhat simplified description follows.

The output of the auditory model is a neurogram (or *neural spectrogram*) depicting the time of firing of a set of 128 simulated auditory nerve fibers in response to each stimulus applied at time $t = 0$ at a simulated sound pressure level of 65 dB. Spacing of the filters in the frequency dimension, according to the Greenwood (1961) equation, corresponds to equal increments of distance along the basilar membrane. Because of the tonotopic (frequency–place) organization of auditory nerve fibers and the systematic spacing of the filters across the 0–5 kHz frequency range, the neural spectrogram is a very effective time–frequency representation. The high data rate associated with the full representation is dramatically reduced by summing nerve firings (spikes) within time–frequency cells to produce a two-dimensional matrix. Spikes are counted in a (12×16)-bin region stretching from -25 to 95 msec in 10-msec steps in the time dimension and from 1 to 128 in steps of eight in the frequency (fiber CF index) dimension. Thus, the nets have a maximum of 192 inputs. These time limits were chosen to exclude most (but not all) of the prestimulus spontaneous activity and the region where responses were expected to be entirely characteristic of the vowel. The impact on synthetic CP of the number and placement of these time–frequency cells has not yet been investigated systematically, just because the initial scheme that we tried worked so well. Some prior thought was given to the resolutions chosen. The 10-msec width of the time bin corresponds approximately to one pitch period. The grouping into eight contiguous filters, in conjunction with equi-spacing according to the Greenwood equation, corresponds to a cell width that is an approximately constant fraction (about 0.7) of the critical bandwidth.

Since the auditory preprocessor is stochastic in nature (because of its simulation of mechanical-to-neural transduction in the hair cells of the cochlea), repetitions of the same input stimulus produce statistically different outputs. This fact is very convenient: It means that a sufficiently large data set for training the neural models can be generated simply by running the preprocessor repeatedly with the same input. In this work, 50 repetitions were used for each of the three (bilabial, velar, and alveolar) series, to

produce neural spectrograms for training and testing the nets.

Brain-State-in-a-Box Model

There was a distinct net for each of the (bilabial, alveolar, and velar) stimulus series. The input data were first reduced to (approximately) zero-mean bipolar patterns by subtracting 5 from each value. This was sufficient to ensure that negative saturating states were appropriately used in forming attractors, in addition to positive saturating states. Initially, simulations used all 192 inputs. A possible problem was anticipated as follows. The number of potential attractor states (corners of the box) in the BSB model grows exponentially with the number of inputs: In this case, we have 2^{192} potential attractors. Clearly, with such a large number, the vast majority of states must remain unlabeled. This will only be a problem, however, if a test input is actually in the region of attraction of such an unlabeled (rubbish) state. In the event, this did not happen. However, training was still unsuccessful in that the different endpoint stimuli (canonical voiced or 0-msec VOT, and canonical unvoiced or 80-msec VOT) were attracted to the same stable states: There was no differentiation between the different endpoints. This was taken as an indication that the full 192-value patterns were more similar to one another than they were different.

In view of this, the *most important* time–frequency cells were identified by averaging the endpoint responses and taking their difference. The N cells with the largest associated absolute values were then chosen to form the inputs to an N-input, N-unit BSB net. This is a form of orthogonalization. Ideally, this kind of preanalysis is best avoided: The neural model ought to be powerful enough in its own right to discover the important inputs. Preliminary testing indicated that results were not especially sensitive to the precise value of N, provided it was in the range somewhere between about 10 and 40. A value of 20 was therefore chosen. These 20 most important time–frequency cells are located around the low-frequency region (corresponding to 200–900 Hz) just after acoustic stimulus onset, where voicing activity varies maximally as VOT varies. The precise time location of this region shifts in the three nets (bilabial, alveolar, and velar) in the same way as does the boundary point. The nets were then trained on the 0- and 80-msec endpoints, and generalization was tested on the full range of stimuli, including the (unseen) intermediate (10–70 msec) stimuli.

Because of the relatively large number (100) of training patterns contributing to the feedback matrix (Equation 2) and, hence, to the weights, it was necessary to increase the neuron saturation limit markedly (to $C = \pm 20{,}000$). The external scale factor was set at $\alpha = 0.1$, and the internal scale factor at $\beta = 0.05$. These values were arrived at by trial and error; network behavior was not especially sensitive to the precise settings. Again, self-connections between neurons were allowed. It was found that the 0-msec (voiced) training patterns were always assigned different corners from the 80-msec (unvoiced) patterns. During generalization testing, no rubbish states were encountered: Convergence was always to a labeled attractor. Moreover, the activation vector (after convergence) for the 0-msec stimuli was found to be the same after training for all three series (i.e., the voiced stimuli all shared the same attractors, irrespective of place of articulation). The same was true of the 80-msec (unvoiced) endpoint stimuli. (This would, of course, be a problem if the task of the net were to identify the place of articulation, rather than the presence/absence of voicing.)

Figure 5A shows the identification function obtained by plotting the proportion of the 50 presentations which converged to a state labeled *voiced* for each of the three series; Figure 5B shows the one-step discrimination function (averaged over 1,000 presentations) obtained using the procedure of Anderson, Silverstein, et al. (1977), as was described in the Brain-State-in-a-Box subsection above. The results are clear and unequivocal: Classical categorization was observed with a steep labeling curve and an ABX discrimination peak at the category boundary. Although the labeling curve was rather too steep and the actual boundary values obtained were slightly low (by about 5 or 10 msec), the shift with place of articulation was qualitatively correct. The finding of correct order of boundary placement was very consistent across replications with different scale factors. We take this to be an indication of its significance. With these more realistic input patterns, there was no need to add noise, as there was in the case of the artificial (vectors) input.

Figure 6A shows the alveolar labeling curve from Figure 5 plotted together with the Kuhl and Miller (1978) human and chinchilla data. This confirms that the BSB model's synthetic identification functions were a reasonable, but not exact, replication of the human and animal data. It was not possible to apply probit analysis to determine the phonetic boundary for the (alveolar) BSB model, because there was only a single point that was neither 100% or 0%. Obviously, the boundary was somewhere between 20 and 30 msec. Also, the synthetic function was closer to the chinchilla data than to the human data. The root-mean square (RMS) difference between the BSB function and the animal data was 19.2 percentage points, whereas the corresponding figure for the human data was 27.8 percentage points. (The RMS difference between Kuhl & Miller's animal and human alveolar labeling data was 10.5 percentage points.) The findings were similar for the bilabial and velar stimuli.

Figure 7 shows the obtained one-step discrimination function for the alveolar series and that predicted on the basis of Equation 1. They are essentially identical, differing only because the obtained function was contaminated by the sampling statistics of the guessing process.

Back-Propagation Network

In light of the foregoing review, there are (at least) two ways that a model of synthetic CP based on back-propagation training of a feedforward net can be produced: (1) As with the BSB model (and paralleling the animal

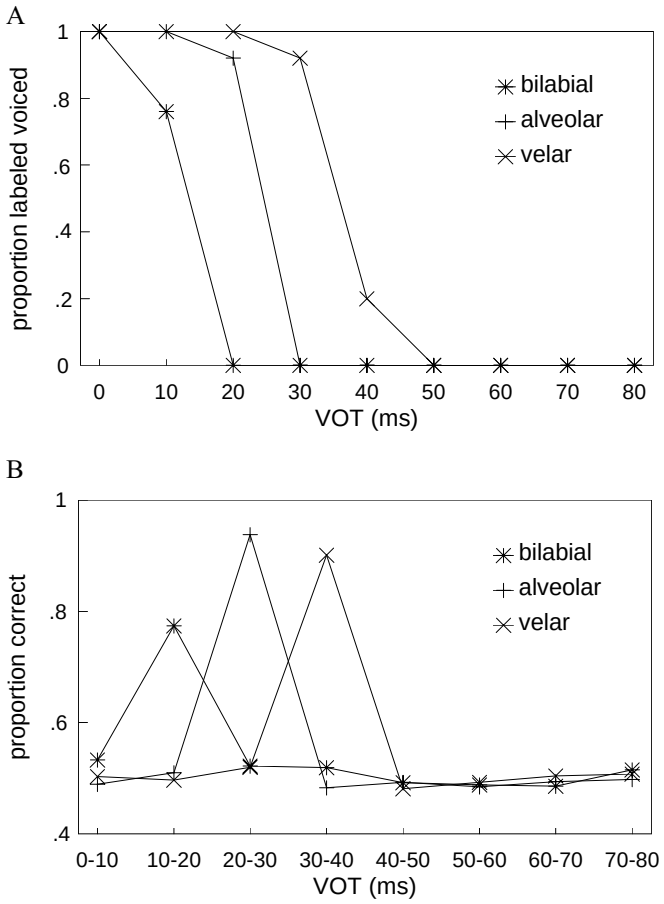

Figure 5. Categorical perception of voice onset time (VOT) in the brain-state-in-a-box model. (A) Labeling functions for bilabial, alveolar, and velar series. Each point is an average of 50 presentations. (B) One-step ABX discrimination functions. Each point is an average of 1,000 presentations.

experiments of Kuhl & Miller, 1978), the net is trained on the 0- and 80-msec endpoints, and generalization is then tested, using the full range of VOT stimuli; and (2) using the auto-association paradigm of Harnad et al. (1991, 1995), hidden-unit representations resulting from pre- and postcategorization training are compared.

In this work, we have adopted the first approach, mostly because we have both psychophysical and synthetic (BSB model) data against which to assess our simulation. This was not the case for Harnad et al.'s (1991, 1995) artificial data, which accordingly required some other reference for comparison.

Initially, a separate MLP was constructed for each of the three (bilabial, alveolar, and velar) series. Each of the three nets had 192 input units, a number (n) of hidden units, and a single output unit (with sigmoidal activation function) to act as a voiced/unvoiced detector. Each net was trained on 50 repetitions (100 training patterns in all) of the endpoint stimuli. The number n of hidden units turned out not to be at all important. In fact, we did not need hidden units at all. Damper, Gunn, and Gore (2000) show that synthetic CP of the VOT continuum is exhibited by single-layer perceptrons, and they exploit this fact in identifying the neural correlates of CP at the auditory nerve level. We used $n = 2$ in the following. Suitable training parameters (arrived at by trial and error) were as follows: learning rate, $\eta = .005$; momentum = .9; weight range = 0.05; error criterion, $\epsilon = 0.25$. The ϵ error criterion was determined by allowing an average error of .05 (or .0025 when squared) for each of the 100 training patterns.

Table 1 shows the result of training the bilabial net 10 times from different initial weight settings. As can be seen, the net trained to the .25 error criterion very easily—typically, in about 50 epochs. The strong tendency, especially for those cases in which the criterion was reached quickly, was to encode the 0-msec endpoint with hidden unit activations of $h_1h_2 = 01$ and the 80-msec endpoint with $h_1h_2 = 10$. (Of course, h_1 and h_2 were never *exactly* 0 or 1 but, more typically, something like .05 or .95.) On only one exceptional occasion (when training required 230 epochs) was a hidden-unit coding arrived at for which

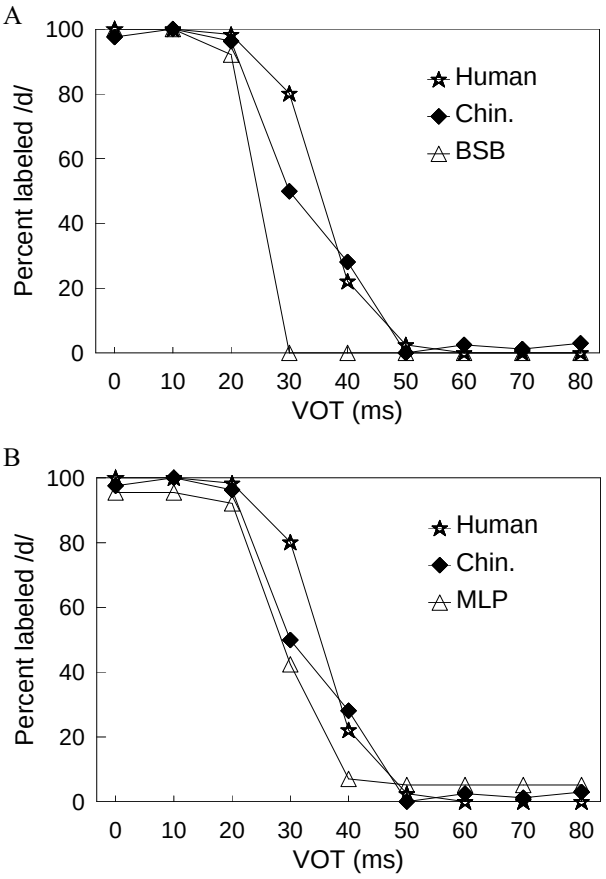

Figure 6. Composite labeling functions for the alveolar series for humans, chinchillas, and neural networks. The human and animal data are taken from Kuhl and Miller (1978, their Figure 3). (A) Brain-state-in-a-box neural model. (B) Multilayer perceptron. VOT, voice onset time.

h_1 and h_2 for the different endpoints were not *both* different. Similar results were obtained for the two other (alveolar and velar) nets, except that the alveolar net was rather more inclined to discover the $h_1h_2 = 00/11$ coding. Seven weight sets were selected for subsequent testing—namely, those obtained in fewer than 100 training epochs.

Figure 8A shows typical labeling functions (from the seven of each) obtained by averaging output activations over the 50 stimulus presentations at each VOT value for the three nets. This averaging procedure avoids any necessity to set arbitrary decision threshold(s) to determine whether the net's output is a *voiced* or an *unvoiced* label: We simply interpret the average as the proportion labeled *voiced*. The reader might question the validity of the averaging operation, since a real listener would obviously not have available in auditory memory a statistical sample of responses from which the average could be computed. Taking the average, however, is a simple and convenient procedure that may not be too different from the kind of similarity measure that could conceivably be computed from a set of prototypes stored in long-term memory. (In any event, it parallels what Anderson, Silverstein, et al., 1977, did in their simulation.) Again, classical CP was observed in all seven cases, with a steep labeling function and separation of the three curves according to place of articulation.

The boundary values found by probit analysis (Finney, 1975), averaged across the seven repetitions, were 20.9, 32.8, and 41.6 msec for the bilabial, alveolar, and velar stimuli, respectively. These are in excellent agreement with the literature (see Table 2), at least in the case of the alveolar and velar stimuli. The labeling curve for the bilabial series in Figure 8 is not as good as those for the alveolar and velar stimuli, with the average activation being rather too low at 20-msec VOT and somewhat too high for VOTs greater than 30 msec. Damper et al. (1990) deal at some length with a possible reason for this, which has to do with the details of the synthesis strategy. To use their description, the bilabial stimuli are *pathological*. It is interesting that the BSB model also seems to be sensitive to this pathology, producing too small a VOT value for the bilabial category boundary (see Figure 8A). The effect was also found (unpublished results) for a competitive-learning net trained with the Rumelhart and Zipser (1985) algorithm. The boundary movement with place of articulation is an emergent property of the nets (see the detailed comments in the Discussion section below). There is no sense in which the nets are explicitly trained to separate the boundaries in this way.

Figure 6B above, shows the typical synthetic identification curve of Figure 8A for the alveolar MLP, as compared with the Kuhl and Miller (1978) human and chinchilla data. It is apparent that the MLP is a rather better model of labeling behavior than is the BSB. By probit analysis, the alveolar boundary was at 32.7 msec (as compared with 33.3 msec for chinchillas and 35.2 msec for humans), which was little different from the average value of 32.8 msec for the seven repetitions. The RMS difference between the MLP function and the animal data was 8.1 percentage points, whereas the difference for the human data is 14.2 percentage points. These figures were about half those for the BSB model. Again, the findings were similar for the bilabial and velar stimuli.

Consider now the discrimination functions for the MLPs. Unlike the Anderson, Silverstein, et al. (1977) simulation, which produced a discrete code, the MLPs produce a continuous value in the range (0,1), because of the sigmoidal activation function. This means that we are not forced to use covert labeling as a basis for the discrimination. We have simulated a one-step ABX experiment, using the Macmillan et al. (1977) model, in which there is first a 2IFC subtask to determine the order of the standards, ⟨AB⟩ or ⟨BA⟩, followed by a yes–no subtask. The A and B standards were selected at random from adjacent classes—that is, from the sets of 50 responses at VOT values differing by 10 msec. The X focus was chosen at random, with equal probability of .5, from one of these

Table 1
Results of Training the 192–2–1 Bilabial Multilayer Perceptron to an Error Criterion of $\epsilon = .25$, Starting From 10 Different Initial Weight Settings

	h_1h_2 Coding		Both
Iterations	0 msec	80 msec	Different
48	01	10	Y
49	01	10	Y
51	01	10	Y
51	10	01	Y
54	01	10	Y
56	01	10	Y
56	10	01	Y
107	00	11	Y
125	11	00	Y
230	01	00	N

two classes. Output activations were then obtained from each of the inputs A, B, and X. Because of the perfect "memory" of the computer simulation, it is possible to collapse the two subtasks into one. Let the absolute difference in activation between the X and the A inputs be |X − A|; similarly, for the B input, the difference is |X − B|. The classification rule is, then,

$$\text{X is } \begin{cases} \text{A} & \text{if } \left|\text{X}-\text{A}\right| < \left|\text{X}-\text{B}\right| \\ \text{B} & \text{otherwise.} \end{cases} \qquad (4)$$

Finally, this classification is scored as either correct or incorrect.

We found, however, that |X − A| and |X − B| were occasionally almost indistinguishable in our simulations, in that they differed only in the fourth or fifth decimal place. In terms of the Macmillan et al. (1977) model, this means that a real listener in the 2IFC subtask would probably have yielded identical (AA or BB) outcomes, which are inappropriate for the yes–no subtask. To avoid making the simulation too sensitive to round-off errors and to simulate the nonideality of real listeners, we therefore introduced a *guessing threshold*, *g*. According to this, X was only classified by the rule of Equation 4 above if

$$\left|\,|\,\text{X}-\text{A}\,|-|\,\text{X}-\text{B}\,|\,\right| > g.$$

If this inequality was not satisfied, the classification of X was guessed with an equal probability of .5 for each class. The results were not especially sensitive to the actual value of *g*.

Figure 8B shows the discrimination functions obtained from such a simulated ABX experiment. There were 500 ⟨ABA⟩ presentations and 500 ⟨ABB⟩ presentations, 1,000 in all. Again taking advantage of a computational shortcut, there were no presentations of the BA standard, on the grounds that the simulation had perfect memory, so that symmetry was ensured and this condition would be practically indistinguishable from the ⟨AB⟩ standard. (In the real situation, of course, memory for the standard presented in the second interval of the ⟨AB⟩ or ⟨BA⟩ dyad would generally be better than that for the standard in the first interval.) The guessing threshold, *g*, was 0.001. There were clear peaks at the phoneme boundary, and the movement of these peaks with the place of articulation was qualitatively correct. Paralleling the less steep (and more psychophysically reasonable) labeling functions, the discrimination peaks were not as sharp as those for the BSB model. They were closer to those typically obtained from real listeners.

Figure 9 shows the discrimination curve obtained from the simulation described above (for the alveolar series) and that predicted from labeling, using the Haskins formula. Similar results were obtained for the other VOT series. The back-propagation model (unlike the BSB model) convincingly reproduces the important effect

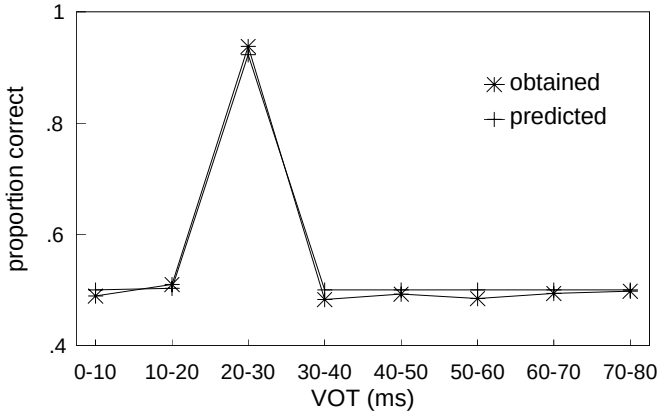

Figure 7. Categorical perception of voice onset time (VOT) in the brain-state-in-a-box model: Obtained ABX discrimination function for the alveolar series and that predicted on the basis of the corresponding labeling function, using the Haskins formula. The close correspondence reflects the fact that the neural model converts the input to a discrete label, so that its discrimination function is a direct implementation of the Haskins formula.

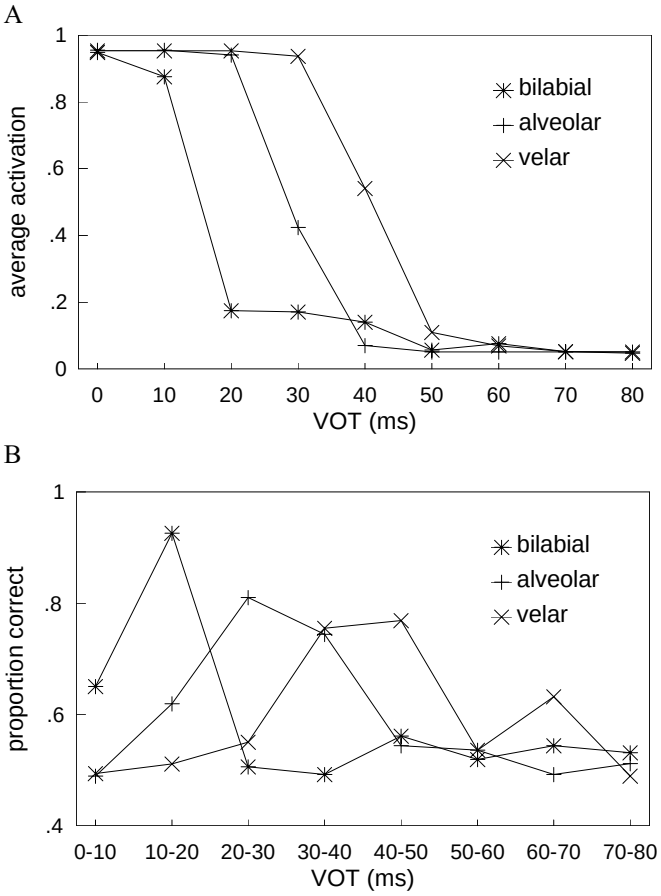

Figure 8. Categorical perception of voice onset time (VOT) by multilayer perceptrons with two hidden units. (A) Labeling functions in terms of average activation. Each function (bilabial, alveolar, and velar series) is obtained from a different net, and each point is an average of 50 presentations. (B) Discrimination functions from a simulated one-step ABX task. Each point is an average of 1,000 presentations: 500 of ⟨ABA⟩ and 500 of ⟨ABB⟩. The guessing threshold *g* was 0.001.

whereby observed discrimination in psychophysical tests exceeds that predicted from Equation 1. This can be interpreted as evidence for the importance of continuous representation(s) in CP.

We have so far been unable to train a *single* net to label these data properly. Although a 192–2–1 net trains easily on all six endpoint stimuli, it will not generalize in such a way as to put the boundary in the correct location for each of the three series of inputs. We are conducting studies to see whether this problem can be overcome by using more hidden units and/or a different output coding.

Figure 10 shows the standard deviation of the activation versus VOT. As one would expect, this tends to peak at the category boundary (although this is clearer, in this particular figure, for the alveolar and velar series than for the bilabial stimuli), which is consistent with our remarks above about this series being pathological. Indeed, the standard deviation could be taken as a credible predictor of reaction time in human psychophysical experiments.

SYNTHETIC CATEGORICAL PERCEPTION AND SIGNAL DETECTION THEORY

In view of the importance of Macmillan et al.'s (1977) contribution to the field, it seems clear that synthetic CP should be assessed by using the techniques they have pioneered. Yet, apart from the work of Eijkman (1992), we know of no other suggestion in the literature to the effect that the methods of psychophysics in general, and of signal detection analysis in particular, are relevant to the evaluation of neural net models. (Even then, Eijkman does little more than advance d' as a useful measure of separation in a pattern-recognition context: His main concern is with his "black box image" technique for visualizing the internal structure of a net.) In this section, we consider the relation between labeling/identification and discrimination from the perspective of SDT.

In the standard yes–no detection task, hits are *yes* responses to signal presentations, and false alarms are *yes*

Table 2
Summary Phonetic Boundary Data for Humans, Chinchillas, and the Multilayer Perceptron (MLP) Neural Model

Boundary Values	Bilabial (msec)	Alveolar (msec)	Velar (msec)
Human			
Pooled	26.8	35.2	42.3
Range	21.3–29.5	29.9–42.0	37.2–47.5
Chinchilla			
Pooled	23.3	33.3	42.5
Range	21.3–24.5	26.7–36.0	41.0–43.7
MLP			
Averaged	20.9	32.8	41.6
Range	18.6–23.4	30.7–35.1	39.8–45.0

Note—Human and chinchilla data are from Kuhl and Miller (1978), and *pooled* means that the identification data were aggregated before fitting a sigmoid and taking its 50% voice onset time (VOT) value. There were 4 human listeners and 4 chinchillas, except for the bilabial and velar conditions, in which only two chinchillas participated. Figures for the MLP are for seven repetitions of training, starting from different random initial weight sets, and *averaged* means that the 50% VOT boundary values were obtained individually and then averaged.

responses to signal-plus-noise presentations. In the ABX discrimination paradigm, it is arbitrary whether ⟨ABA⟩ and ⟨BAB⟩ are taken to correspond to signal presentations and ⟨ABB⟩ and ⟨BAA⟩ are taken to correspond to signal-plus-noise or vice versa. We adopt the former convention, as in Table 3, where *Response 1* means that the subject nominated the first interval as containing the standard corresponding to X and *Response 2* means that the second interval was nominated. In the simulations described in the previous section, the perfect memory of the computer simulations means that the order of the A and B standards was irrelevant. Hence, as was previously stated, only the stimulus–response matrix in the top half of the table was collected.

Macmillan et al. (1977) consider the unbiased case of ABX discrimination. This, they say, is "the only one for which a simple expression can be written for the hit and false alarm rates" (p. 459) in the form of their Equation 3:

$$H = P(\text{Response 1} \mid \langle \text{ABA} \rangle) = P(\text{Response 1} \mid \langle \text{BAB} \rangle) = 1 - FA.$$

A d'-like sensitivity measure can now be obtained as

$$\begin{aligned} d'_s &= z(H) - z(FA) \\ &= z(H) - z(1 - H) \\ &= 2z(H), \end{aligned} \tag{5}$$

from the hit rate alone. True d' can then be found from this d'_s, using Table 3 of Kaplan, Macmillan, and Creelman (1978). In the terms of Macmillan et al. (1977), CP requires that this true d' for discrimination shall be equivalent to an *identification distance*, or identification d', obtained by the Braida and Durlach (1972) procedure. This involves subtracting the z-transformed probabilities of assigning presentations to the same category.

In the following, we analyze only the back-propagation neural model of VOT perception. We exclude the BSB model from consideration for two reasons. First, according to the results of the previous section, it produces a less convincing simulation of the psychophysical data. (Recall, also, that it was necessary to orthogonalize the input data for the BSB model, but not for the back-propagation model.) Second, it is inherently unsuitable for the analysis, because its labeling function includes many 1 and 0 points, which are not amenable to transformation to z-scores, since they yield values of $\pm\infty$.

Table 4 shows the (one-step) discrimination d', found by using Equation 5 and Table 3 of Kaplan et al. (1978), and the identification d' obtained from the z-transformed identification proportions for the representative case of the alveolar series. In Figure 11, discrimination d' is plotted against identification d'. According to the Macmillan et al. (1977) characterization of CP, these two measures should be equivalent. That is, the points of Figure 11 should be distributed about the unity-slope straight line. Clearly, discrimination d' exceeds identification d', as is so often found: The slope of the best-fit line is actually 1.899 (unaffected by forcing the regression through the origin). However, on the basis of a paired t test ($t = 1.2926$ with $\nu = 7$ degrees of freedom, two-tailed test), we reject the hypothesis that these two sensitivity measures are equivalent ($p \sim .2$). Similar findings hold for the bilabial and velar series. As in other studies, the two measures are highly correlated. Regression analysis (alveolar series) yields $r = .8740$ for the case in which (on the basis of assumed equality) the best-fit line is forced to pass through the origin, and $r = .8747$ when it is not ($p < .1$). Hence, we conclude that discrimination performance is correlated with but somewhat higher than identification performance, as in the case of human observers (e.g., Macmillan, 1987, p. 59).

DISCUSSION: IMPLICATIONS OF SYNTHETIC CATEGORICAL PERCEPTION

Thus far, we have emphasized CP as an emergent property of learning systems in general, arguing that it is not a *special* mode of perception. In this section, we aim to make these claims more concrete. If CP is indeed an emergent and general property of learning systems, one might ask, Why are strong CP effects not *always* found (e.g., in vowel discrimination or long-range intensity judgments)? Rather, degree of CP is observed to vary with the nature of the stimuli, the psychophysical procedure, the experience of the subjects, and so on. To answer this key question, we draw a sharp distinction between the two essentially different kinds of synthetic CP that have been explored in this paper.

Consider first the work using artificial or novel stimuli, such as that of Anderson, Silverstein, et al. (1977), Harnad et al. (1991, 1995), and Goldstone et al. (1996). In these cases, the category boundary is placed either (1) at a point

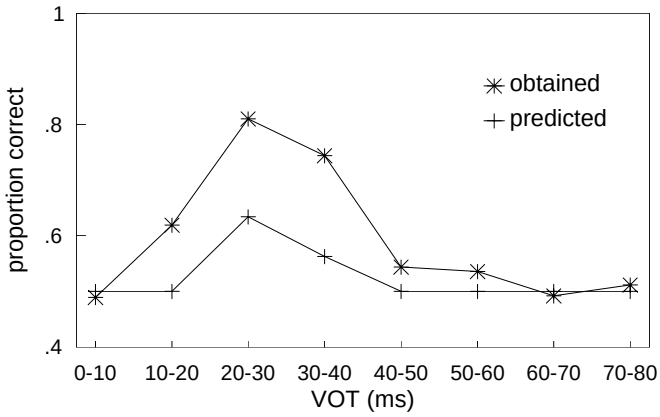

Figure 9. Categorical perception of voice onset time (VOT) by a multilayer perception with two hidden units: the one-step ABX discrimination function obtained for the alveolar series and that predicted on the basis of the corresponding labeling function, using the Haskins formula. Here, the obtained discrimination is better than that predicted (as can be seen in the psychophysical results), reflecting the fact that the output of the neural model is continuous, rather than discrete.

predetermined by the labels supplied during (supervised) training, if training was on the complete continuum, or (2) at the center of the continuum, if training was on the endpoints of the continuum only. This corresponds to the situation in which real subjects would not already possess internal labels, anchors, or prototypes and so would only display CP as a result of training and experience. Hence, this type of synthetic CP can only be a reasonable model of learned categorization—such as that found in the categorization training experiments of Goldstone et al., as well as those of Goldstone (1994), Beale and Keil (1995), Pevtzow and Harnad (1997), Livingstone, Andrews, and Harnad (1998), Goldstone (1998), and Stevenage (1998)—rather than of innate categorization. By contrast, in the work on stop consonants, the nets are trained on endpoint stimuli from a VOT continuum, and generalization of the learning to intermediate stimuli is tested. In this case, the synthetic listener places the category boundaries of the three (bilabial, alveolar, and velar) series in a way that predicts the psychophysical results from real listeners.

A very revealing finding in this regard is that the BSB simulation categorizes the artificial (vectors) continuum and the VOT stimuli very differently, even though training is on endpoints in both cases. Figure 3B shows that the category boundary is precisely at midrange, between Stimuli 7 and 8. This is hardly surprising: It is difficult to see what else the net might do to dichotomize the data, other than bisect the continuum at its midpoint. On the other hand, Figure 5A shows that the BSB nets position

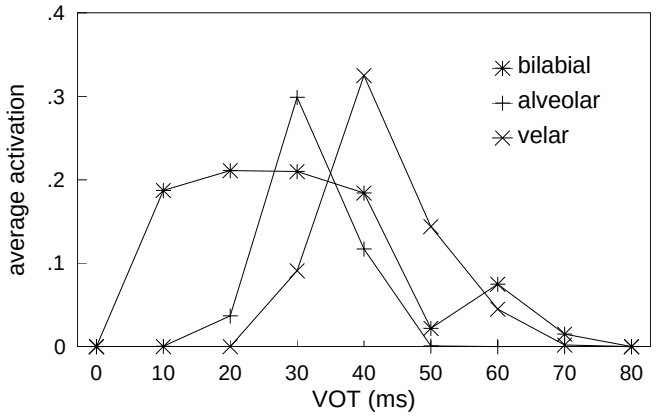

Figure 10. Standard deviation of the multilayer perceptron's output activation. Each point is an average of 50 presentations. VOT, voice onset time.

Table 3
Stimulus–Response Matrix for the ABX Discrimination Paradigm

Presentation	Response 1	Response 2
⟨ABA⟩	hit	miss
⟨ABB⟩	false alarm	correct rejection
⟨BAA⟩	false alarm	correct rejection
⟨BAB⟩	hit	miss

their phonetic boundaries in such a way as to segregate the VOT series by place of articulation in the same way human and animal listeners do. This striking difference in network behavior can only be explained credibly by the different input continua: It is unlikely that it could have arisen through essentially trivial differences in parameters, such as the different numbers of inputs in the two cases (16 as opposed to 20). Thus, we infer that there is some property of the input continua in the simulation of VOT perception that is not shared by the much simpler artificial/vectors continuum. Hence, we do not expect to observe strong CP effects in all cases of generalization testing but only when the stimulus continuum (appropriately encoded for presentation to the net) has some special properties. That is, the potential for categorization must be somehow implicit in the physical stimulus continuum and its encoding schema. Because they are embodied in software, connectionist models can be systematically manipulated to discover their operational principles. Thus, means are available to discover just what these special properties might be. Damper et al. (2000) show that each of the three (bilabial, alveolar, and velar) nets has its strongest connections to different areas of the neurogram and that these differences predict the shift of boundary with place of articulation. We infer that what is supposed to be a continuum actually is not. There are discontinuities (systematically dependent on place of articulation) in the Abramson and Lisker (1970) stimuli themselves, and this is the sense in which the potential for categorization exists. In other words, what is supposed to be a unidimensional continuum (so that only VOT and features perfectly correlated with it vary) is actually multidimensional.

Of course, in the case of the VOT stimuli, the inputs to the BSB net have been subjected to sophisticated preprocessing by the auditory model of Pont and Damper (1991). The role this (simulated) auditory processing plays in the observed categorization behavior is currently being investigated. Early indications (Damper, 1998) are that the auditory preprocessing is vital to realistic simulation of VOT perception in that "the front-end processor is not essential to category formation but plays an important part in the boundary-movement phenomenon, by emphasizing . . . parts of the time–frequency regions of the speech signal" (p. 2196).

We are now in a position to refine our notion of *emergent* functionality. The concept of emergence has grown in popularity in cognitive science and neural computing in recent years and is starting to influence models of speech categorization (e.g., Guenter & Gjaja, 1996; Lacerda, 1998). The term does not easily admit of a precise definition (see, e.g., Holland, 1998, p. 3), but Steels (1991) writes: "Emergent functionality means that a function is not achieved directly by a component or a hierarchical system of components, but indirectly by the interaction of more primitive components among themselves *and with the world*" (p. 451, italics added). This seems to capture rather well what is going on here: The primitive components are the units of the neural net(s) that interact with "the world" in the form of external stimuli, with sensory transduction and/or early perceptual processing mediating between them. The interaction "with the world" is particularly important. The potential for categorization must exist implicitly in the sensory continuum. So, in what sense is CP not *special*? From the work described in this paper, it is apparent that we do not need specialized processing apparatus, as is posited in motor theory. Rather, provided the sensory continua are of the right form, any general learning system operating on broadly neural principles ought to exhibit the essentials of CP.

Finally, we note that the TRACE model apparently acts like the BSB model, simply placing the phonetic boundary between /g/ and /k/ at midrange. This could be either because the stylized inputs to the net (interpolated VOT and $F1$ onset frequency) are not good counterparts to the

Table 4
Signal Detection Theory Analysis of Synthetic Categorical Perception by the Back-Propagation Neural Model: Alveolar Series

VOT (msec)	Identification Proportion	Discrimination				
		z(ident)	Identification d'	Hit Rate (H)	$d's = 2z(H)$	Discrimination d'
0	0.955	–	–	–	–	–
10	0.955	1.6950	0.0095	0.4840	−0.0803	−0.3800
20	0.941	1.6855	1.7046	0.8280	1.8926	2.2426
30	0.424	−0.0191	1.4567	0.9720	3.8220	3.9120
40	0.070	−1.4758	0.1593	0.5260	0.1305	0.4810
50	0.051	−1.6351	0.0	0.4220	−0.3936	−0.8736
60	0.051	−1.6351	0.0	0.4500	−0.2513	0.6826
70	0.051	−1.6351	0.0	0.5180	0.0903	0.4006
80	0.051	−1.6351	0.0	0.5040	0.0201	0.1904

Note—VOT, voice onset time.

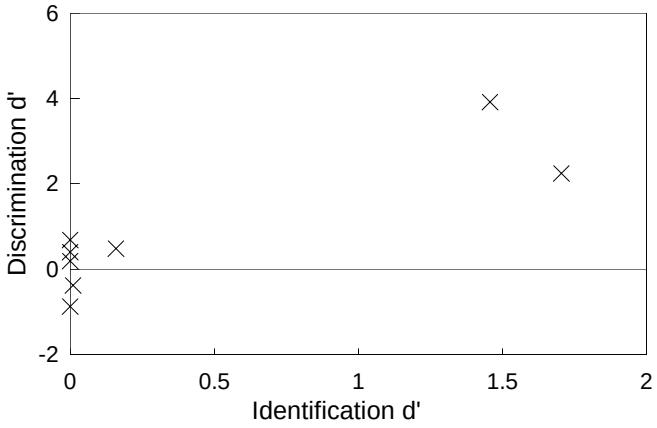

Figure 11. Discrimination *d′* versus identification *d′* for the back-propagation neural model and the alveolar voice onset time series. Categorical perception according to Macmillan, Kaplan, and Creelman (1977) requires that these two sensitivity measures are equivalent (i.e., in the ideal case, all the points should lie on the unity-slope line).

Abramson and Lisker (1970) stimuli or, more likely, because there has been no (simulated) auditory preprocessing of these patterns. Further work is necessary to distinguish between these two possibilities.

CONCLUSIONS AND FUTURE WORK

Neural nets provide an underexplored yet revealing way of studying CP. We have shown that a variety of neural models is capable of replicating classical CP, with the point of maximal ambiguity of the steep labeling function and a sharp peak of the discrimination function coinciding at the category boundary. Given the ubiquitous way that CP arises in network performance, we believe that the effect is very basic to how such nets (and other adaptive systems, such as human and animal subjects) learn. Focusing on the special case of speech CP with initial stop consonants, we have also demonstrated the shift of phoneme boundary with place of articulation for the voicing continuum by using two different nets—the historically important BSB model of Anderson, Silverstein, et al. (1977) and a more recent MLP (back-propagation) model. The most convincing demonstration of synthetic CP to date is that by the MLP model. The categorization behavior is an emergent property of the simulations: There is no sense in which it is programmed into the model or results from parameter adjustment in a curve-fitting sense. The back-propagation net also replicates the frequently documented effect whereby observed discrimination performance exceeds that predicted from labeling on the assumption that only discrete information about the labels is available (the so-called Haskins model). It does so by retaining continuous information after the stage of sensory processing. That is, the categorization occurs at the later, decision stage.

Early studies of CP considered the effect in terms of low-threshold theories, which assume a mapping of sensory stimuli to discrete internal states. In 1977, Macmillan et al. made an important advance by applying to CP the more modern methods of SDT, which assume a continuous internal representation. They suggested that CP should be characterized by the equivalence of identification and discrimination sensitivity, both measured with *d′*. Our back-propagation simulations fail to satisfy this definition, in that identification *d′* is statistically different from discrimination *d′*, although the two are correlated.

The Macmillan et al. (1977) paper is now 20 years old. So, despite its pioneering nature, it is obviously not the last word on CP. Indeed, since 1977, Macmillan has retreated somewhat from the position that equivalence of discrimination and identification should be taken to be the defining characteristic of CP (see Macmillan, Goldberg, & Braida, 1988). In 1987, he writes: "it is clear that few if any dimensions have this property" (p. 78). Nonetheless, "relations between tasks *can* provide useful information about the manner in which stimuli are processed, however such processing is named." In the present work, we have shown that the relation between identification and discrimination performance for a simple neural net simulation of VOT perception closely parallels that seen for real subjects. Since the simulation is faithful in other respects too, it can (and should) be taken seriously as a model of CP. Because of the simplicity of the model and the way that category boundary phenomena arise quite naturally during learning, we conclude that CP is not a special mode of perception. Rather, it is an emergent property of learning systems in general and of their interaction with the stimulus continuum,[11] mediated by sensory transduction and/or early perceptual processing.

The assumptions underlying our models are similar to those of Nearey (1997, note 1), who presupposes, as do we, "the segmentation of the input signals . . . before they are presented to a perceptual model." In addition, because of their inability to handle temporal sequences of inputs,

the models are assumed to have perfect memory (see Port, 1990, for criticisms of improper treatment of time in connectionist models). In the case of VOT perception, the reduced (192-value) neural spectrogram is available as a (conveniently presegmented) static input to the backpropagation net. Apart from an implicit *time as space* representation, there is no explicit representation of relational time. Precise time representation seems unnecessary for the credible modeling of VOT perception, since the spike-counting procedure (which reduces the neurogram to a 192-component vector for presentation to the MLP) effectively obscures this. The BSB model was unable to distinguish the voicing contrast in the complete 192-value patterns and was, therefore, given 20 selected input values only, again as a static pattern.

Real listeners obviously hear a sequence of sounds during speech communication, and memory effects are a very important component of perception. Hence, a priority for future work is the addition of recurrent (feedback, as opposed to purely feedforward) connections to the more realistic perceptron model, in the manner of Jordan (1986) or Elman (1990), so as to implement an imperfect memory buffer. Future studies should also address the synthetic categorization of vowel continua and the precise role of preprocessing by the (simulated) auditory periphery. Much could also be learned from studying a real (rather than synthesized) stop consonant continuum, provided sufficient productions could be gathered from the ambiguous region around the phonetic boundary.

Finally, an important question concerns the role of the auditory model vis á vis the input stimuli. What would happen if we were to apply our analyses directly to the input patterns without the (rather complex) auditory model's intervening? It is difficult to do this, because there is only a single synthetic token for each acoustic stimulus. (This is unavoidable, since we are no longer simulating the stochastic process of mechanical-to-neural transduction by the cochlear hair cells.) Hence, there is an extreme paucity of data on which to train the neural network model(s). In an attempt to answer this question indirectly, Damper et al. (2000) replaced the auditory front end with a short-time Fourier analysis and then used a support vector machine to model labeling behavior. This kind of learning machine makes best use of sparse training data. It was found that correct movement of the boundary with place of articulation was abolished, indicating that some aspect or aspects of peripheral auditory function are essential to correct simulation of categorization behavior. To confirm that this was not an artifact of having only a single training token per class, perceptrons were trained on single, *averaged* neurograms, whereupon appropriate categorization behavior was maintained (Damper, 1998), indicating that information about the statistical distribution of training data is not essential to the simulation and that the extreme sparsity of the training data need not be fatal. In future work, we intend to confirm these preliminary findings in two ways. First, we will use a database of real speech, so that multiple training tokens whose statistics reflect natural variability in production will be available for training. Second, we will employ a variety of simplified front-end analyses to determine those aspects of the peripheral auditory transformation that are essential to simulating boundary movement with place of articulation.

REFERENCES

Abramson, A., & Lisker, L. (1970). Discrimination along the voicing continuum: Cross-language tests. In *Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967* (pp. 569-573). Prague: Academia.

Amari, S., & Takeuchi, A. (1978). A mathematical theory on formation of category detecting neurons. *Biological Cybernetics*, **29**, 127-136.

Anderson, J. A. (1968). A memory storage model utilizing spatial correlation functions. *Kybernetik*, **5**, 113-119.

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 27-90). Hillsdale, NJ: Erlbaum.

Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.

Anderson, J. A., & Silverstein, J. W. (1978). Reply to Grossberg. *Psychological Review*, **85**, 597-603.

Anderson, J. A., Silverstein, J. W., & Ritz, S. A. (1977). Vowel preprocessing with a neurally based model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 265-269). New York: IEEE.

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

Arbib, M. A. (1995). *Handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53-58.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, **57**, 217-239.

Bégin, J., & Proulx, R. (1996). Categorization in unsupervised neural networks: The Eidos model. *IEEE Transactions on Neural Networks*, **7**, 147-154.

Bornstein, M. H. (1987). Perceptual categories in vision and audition. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 287-300). Cambridge: Cambridge University Press.

Bourland, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291-294.

Brady, S. A., & Darwin, C. J. (1978). Range effects in the perception of voicing. *Journal of the Acoustical Society of America*, **63**, 1556-1558.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, **51**, 483-502.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991a). ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, **4**, 493-504.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991b). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759-771.

Damper, R. I. (1998). Auditory representations of speech sounds in a neural model: The role of peripheral processing. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 2196-2201). New York: IEEE.

Damper, R. I., Gunn, S. R., & Gore, M. O. (2000). Extracting phonetic knowledge from learning systems: Perceptrons, support vector machines and linear discriminants. *Applied Intelligence*, **12**, 43-62.

Damper, R. I., Pont, M. J., & Elenius, K. (1990). *Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus* (Tech. Rep. No. STL-QPSR 4/90). Stockholm: Royal Institute of Technology (KTH), Speech Transmission Laboratory. (Also published in W. A. Ainsworth [Ed.)], *Advances in speech,*

hearing and language processing [Vol. 3, Pt. B, pp. 497-546]. Greenwich, CT: JAI Press, 1996)

Diehl, R. L., Elman, J. E., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 599-609.

Diehl, R. L., & Kluender, K. R. (1987). On the categorization of speech sounds. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 226-253). Cambridge: Cambridge University Press.

Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a mind versus modeling the brain: Artificial intelligence back at a branch-point. *Daedalus*, **117**, 15-43.

Eijkman, E. G. J. (1992). Neural nets tested by psychophysical methods. *Neural Networks*, **5**, 153-162.

Elman, J. L. (1979). Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection theory analysis. *Journal of the Acoustical Society of America*, **65**, 190-207.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**, 179-211.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-380). Hillsdale, NJ: Erlbaum.

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, **83**, 1615-1626.

Finney, D. J. (1975). *Probit analysis: A statistical treatment of the sigmoid response curve* (3rd ed.). Cambridge: Cambridge University Press.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language & Speech*, **5**, 171-189.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, **28**, 67-73.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, **29**, 207-214.

Fujisaki, H., & Kawashima, T. (1971). A model of the mechanisms for speech perception: Quantitative analysis of categorical effects in discrimination. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, **30**, 59-68.

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

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

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

Green, D. M., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: Wiley. (Reprint edition published by Peninsula Press, Los Altos, CA, 1988)

Greenwood, D. D. (1961). Critical bandwidth and the frequency coordinates on the basilar membrane. *Journal of the Acoustical Society of America*, **33**, 780-801.

Grossberg, S. (1968a). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences*, **59**, 368-372.

Grossberg, S. (1968b). Some physiological and biological consequences of psychological postulates. *Proceedings of the National Academy of Sciences*, **60**, 758-765.

Grossberg, S. (1969). Embedding fields: A theory of learning with physiological implications. *Journal of Mathematical Psychology*, **6**, 209-239.

Grossberg, S. (1978). Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz and Jones. *Psychological Review*, **85**, 592-596.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 187-294). London: Academic Press.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, **11**, 23-63.

Guenter, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, **100**, 1111-1121.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral & Brain Sciences*, **13**, 471-518.

Harnad, S. (1982). Metaphor and mental duality. In T. Simon & R. Scholes (Eds.), *Language, mind and brain* (pp. 189-211). Hillsdale, NJ: Erlbaum.

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition*. Cambridge: Cambridge University Press.

Harnad, S., Hanson, S. J., & Lubin, J. (1991). Categorical perception and the evolution of supervised learning in neural nets. In D. W. Powers & L. Reeker (Eds.), *Working papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology* (pp. 65-74).

Harnad, S., Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps towards principled integration* (pp. 191-206). London: Academic Press.

Hary, J. M., & Massaro, D. W. (1982). Categorical results do not imply categorical perception. *Perception & Psychophysics*, **32**, 409-418.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception & Performance*, **8**, 68-80.

Hinton, G. E., & Anderson, J. A. (Eds.) (1981). *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Holland, J. H. (1998). *Emergence: From chaos to order*. Reading, MA: Addison-Wesley.

Howell, P., Rosen, S., Laing, H., & Sackin, S. (1992). *The role of F1 transitions in the perception of voicing in initial plosives* (Tech. Rep. No. 6). University College London, Department of Phonetics and Linguistics.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, **97**, 553-562.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 531-546). Hillsdale, NJ: Erlbaum.

Kaplan, H. L., Macmillan, N. A., & Creelman, C. D. (1978). Tables of *d′* for variable-standard discrimination paradigms. *Behavioral Research Methods & Instrumentation*, **10**, 796-813.

Kohonen, T. (1977). *Associative memory: A system theoretic approach*. Berlin: Springer-Verlag.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Kuhl, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355-386). Cambridge: Cambridge University Press.

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., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, **32**, 542-550.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, **73**, 1003-1010.

Lacerda, F. (1998). An exemplar-based account of emergent phonetic categories [Abstract]. *Journal of the Acoustical Society of America*, **103** (5, Pt. 2), 2980.

Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

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

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language & Speech*, **1**, 153-167.

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

Liberman, A. M., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Liberman, A. M., & Mattingly, I. (1989). A specialization for speech perception. *Science*, **243**, 489-494.

Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops. *Word*, **20**, 384-422.

Lisker, L., & Abramson, A. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967* (pp. 563-567). Prague: Academia.

Livingstone, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **123**, 178-200.

Lloyd, D. (1989). *Simple minds*. Cambridge, MA: MIT Press, Bradford Books.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America*, **103**, 3648-3655.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, **70**, 61-79.

Macmillan, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 53-85). Cambridge: Cambridge University Press.

Macmillan, N. A., Braida, L. D., & Goldberg, R. F. (1987). Central and peripheral effects in the perception of speech and nonspeech sounds. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 28-45). Dordrecht: Nijhoff.

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

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262-1280.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, **84**, 452-471.

Massaro, D. W. (1987a). Categorical partition: A fuzzy logical model of categorical behavior. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 254-283). Cambridge: Cambridge University Press.

Massaro, D. W. (1987b). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. Lass (Ed.), *Speech and language: Vol. 3. Advances in basic research and practice* (pp. 129-165). New York: Academic Press.

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

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

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, **101**, 3241-3254.

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, **89**, 2978-2991.

Pastore, R. E. (1987a). Categorical perception: Some psychophysical models. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 29-52). Cambridge: Cambridge University Press.

Pastore, R. E. (1987b). Possible acoustic bases for the perception of voicing contrasts. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 188-198). Dordrecht: Nijhoff.

Pastore, R. E., Szczesiul, R., Wielgus, V., Nowikas, K., & Logan, R. (1984). Categorical perception, category boundary effects, and continuous perception: A reply to Hary and Massaro. *Perception & Psychophysics*, **35**, 583-585.

Pevtzow, R., & Harnad, S. (1997). Warping similarity space in category learning by human subjects: The role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorisation* (pp. 189-195). Edinburgh: University of Edinburgh, Department of Artificial Intelligence.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328-333.

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

Pont, M. J., & Damper, R. I. (1991). A computational model of afferent neural activity from the cochlea to the dorsal acoustic stria. *Journal of the Acoustical Society of America*, **89**, 1213-1228.

Port, R. F. (1990). Representation and recognition of temporal patterns. *Connection Science*, **2**, 151-176.

Quinlan, P. (1991). *Connectionism and psychology: A psychological perspective on new connectionist research*. Hemel Hempstead, U.K.: Harvester Wheatsheaf.

Repp, B. H. (1984). Categorical perception: Issues, methods and findings. In N. Lass (Ed.), *Speech and language: Vol. 10. Advances in basic research and practice* (pp. 244-335). Orlando, FL: Academic Press.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 129-145.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 89-112). Cambridge: Cambridge University Press.

Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics*, **7**, 393-402.

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, **323**, 533-536.

Rumelhart, D. E., & McClelland, J. L. (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (2 vols.). Cambridge, MA: MIT Press, Bradford Books.

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

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, **15**, 461-508.

Sinex, D. G., & McDonald, L. P. (1988). Average discharge rate representation of voice-onset time in the chinchilla auditory nerve. *Journal of the Acoustical Society of America*, **83**, 1817-1827.

Steels, L. (1991). Towards a theory of emergent functionality. In J.-A. Meyer & S. W. Wilson (Eds.), *From animals to animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 451-461). Cambridge, MA: MIT Press, Bradford Books.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced category perception of identical twin faces. *British Journal of Psychology*, **89**, 39-57.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper,

F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, **77**, 234-239.

Takagi, N. (1995). Signal detection modeling of Japanese listeners' /r/–/l/ labeling behavior in a one-interval identification task. *Journal of the Acoustical Society of America*, **97**, 563-574.

Tijsseling, A., & Harnad, S. (1997). Warping similarity space in category learning by backprop nets. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorisation* (pp. 263-269). Edinburgh: University of Edinburgh, Department of Artificial Intelligence.

Treisman, M., Faulkner, A., Naish, P. L. N., & Rosner, B. S. (1995). Voice-onset time and tone-onset time: The role of criterion-setting mechanisms in categorical perception. *Quarterly Journal of Experimental Psychology*, **48A**, 334-366.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, **91**, 68-111.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, **60**, 1381-1389.

Wood, C. C. (1978). Variations on a theme by Lashley: Lesion experiments with the neural model of Anderson, Silverstein, Ritz and Jones. *Psychological Review*, **85**, 582-591.

NOTES

1. In *identification*, subjects are required to learn (or supply already learned) unique labels to stimuli. In *discrimination*, subjects must (learn to) distinguish between classes of stimuli. Usually, single stimuli are presented in identification, and multiple stimuli in discrimination, but see the recent work of Lotto, Kluender, and Holt (1998), in which "experimental design and stimulus presentation are exactly the same; only the response labels differ" (p. 3649).

2. Dreyfus and Dreyfus (1988) credit James A. Anderson, together with Stephen Grossberg and Tuevo Kohonen, with keeping neural net modeling in artificial intelligence and cognitive science alive "during the dark ages" (their note 8), during which it was generally considered to have been superseded by the physical-symbol system hypothesis of intelligence.

3. As was pointed out to us by Michael Studdert-Kennedy (personal communication, August 7, 1997), the concept of VOT was not formulated until 1964 by Lisker and Abramson. In the 1958 Liberman et al. paper, *F*1-cutback was viewed as a purely acoustic variable: Its origin in VOT was not understood at that time. VOT was originally defined as an articulatory variable—the interval between stop release and the onset of voicing—having multiple acoustic consequences, including the presence/absence of prevoicing, variations in release-burst energy, aspiration duration, and *F*1 onset frequency. In this sense, VOT includes *F*1 onset frequency: The two are (supposedly) perfectly correlated.

4. However, the magnitude of context effects varies greatly with the nature of the stimuli and tends to be negatively correlated with the degree of CP, as was discussed by Repp (1984). Also, the "unresolved difficulty" referred to above arises in part, if not entirely, from the different contexts of the stimuli in typical identification and discrimination tasks (Lotto et al.,1998; Repp, Healy, & Crowder, 1979), at least in the case of vowels.

5. In threshold theories (Luce,1963; Macmillan & Creelman, 1991), a physical continuum is assumed to map to discrete perceptual states, rather than into a perceptual continuum. The threshold is the division between the internal states. In high-threshold theory, the thresholds themselves set the limits to detection, and errors on *noise* trials arise only from guessing. In low-threshold theory, the limit on performance is set by a factor other than the threshold, such as noise (Pastore, 1987a, p. 36).

6. We use angled braces (⟨⟩) to denote an actual presentation dyad or triad (for which X cannot be ambiguous), in contrast to ABX, which is the name of the paradigm.

7. We use the terms *neuron* and *unit* interchangeably.

8. Here, the term *auto-associative* refers to multilayer feedforward nets with hidden units trained to reproduce their inputs as outputs. This is distinct from auto-associative feedback nets like the BSB model, in which there are no hidden units and the *input* and *output* units are physically identical.

9. The reader might object that lines of differing length are not perceived categorically by humans and, so, a network simulation should not show CP either. This is the point of varying the iconicity of the codings: to find the ones that make sense. Literally giving both human observers and networks 12 lines to identify/discriminate will not produce an equivalent task: The coding needs to be varied, to achieve the equivalence.

10. We prefer to reserve the term *attractor* to describe a stable state of a dynamical system. As such, it cannot strictly describe the sort of static input–output relation that Goldstone et al. (1996) clearly have in mind.

11. According to Harnad (1982), "experiential inputs [can] vary continuously along several sensory dimensions, rather than falling neatly into certain prefabricated physical or perceptual categories. . . . [They can be] multidimensional and polysensory . . . such as complex geometric forms, acoustic timbres and sound sequences, complex daily events and sequences of experiences—in fact, any experience that varies along an actual continuum (or a 'virtual' continuum, in virtue of unresolvable information complexity). And this is not yet to have mentioned purely abstract cases, such as the 'space' from which . . . the foregoing list of examples [were extracted]."

(Manuscript received June 6, 1997;
revision accepted for publication April 2, 1999.)