Manuscript to be submitted to *Perception and Psychophysics*, Version 0 (draft): July 4, 1997

The Psychophysics of Synthetic Categorical Perception

R.I. Damper[†*] and S. Harnad[‡]

Cognitive Sciences Centre

University of Southampton

Southampton S017 1BJ

UNITED KINGDOM

([†]also Image, Speech and Intelligent Systems (ISIS) Research Group

Department of Electronics and Computer Science)

([‡]also Department of Psychology)

[*]Center for Spoken Language Understanding

Oregon Graduate Institute of Science and Technology

PO Box 91000

Portland

Oregon 97291-1000

USA

Emails: `rid@ecs.soton.ac.uk` and `harnad@cogsci.soton.ac.uk`

Address for correspondence:

Until 21 July 1997:

Dr. R.I. Damper,
Image, Speech and Intelligent Systems (ISIS) Research Group,
Building 1,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ,
UK.

Tel: +44 1703 594577
FAX: +44 1703 594498
Email: `rid@ecs.soton.ac.uk`

After 21 July 1997:

R.I. Damper, PhD,
Visiting Research Professor,
Department of Computer Science and Engineering,
Oregon Graduate Institute of Science and Technology,
PO Box 91000,
Portland,
OR 97291-1000,
USA.

Tel: +1 503 690-1517
FAX: +1 503 690-1306
Email: `damper@cse.ogi.edu` <u>or</u>
`rid@ecs.soton.ac.uk` (will auto-forward)

Captions to Figures

Captions to Tables

Abstract

Studies of the categorical perception (CP) of sensory continua have a long and rich history in psychophysics. A major development was Macmillan *et al.*'s application in 1977 of signal detection theory to analyze several experimental paradigms, in particular explicating the relation between the psychometric labeling function and discrimination measures. Simultaneously, Anderson *et al.* proposed a neural model for what we will term *synthetic* CP, yet this line of research has been less well explored. In this paper, we assess neural-network models of CP with particular reference to their ability to predict the psychophysical behavior of *real* observers – including the relation between labeling and discrimination. Synthetic categorization of a variety of stimuli, including speech sounds and artificial/novel dimensions, is reviewed and discussed in terms of both classical theories of CP and more recent developments. A variety of neural mechanisms is capable of replicating the essentials of categorical perception, indicating that CP is not a special mode of perception but an emergent property of any sufficiently-powerful general learning system. However, the most convincing replication is from a simulation whose output is continuous rather than discrete.

1 Introduction

Studies of the categorical perception (CP) of sensory continua have a long and rich history: see the volume edited by Harnad (1987) for a comprehensive review up until a decade ago. An important issue has centered on the precise definition of 'categorical perception'. According to the seminal contribution of Macmillan *et al.* [1] (1977, p. 452): "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". The extent to which discrimination can be predicted by 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, so that discrimination is monotonically related to the stimulus – being better at low magnitude. Also, discrimination is much better than identification: observers can only identify a small number of stimuli but can make pairwise discriminations between a much larger number (Miller, 1956). By contrast, CP was originally characterized by an ability to discriminate stimuli which is limited by (and therefore predictable from) identification ability: subjects can only discriminate between categories, not within them. Consequently, the discrimination function is non-monotonic and it peaks at a category 'boundary' (Wood, 1976) – defined as a 50% point on a sharply-rising or falling psychometric labeling function. In his 1984 review, Repp (p. 253) calls this "the essential defining characteristic of categorical perception". However, Macmillan *et al.* (1977) challenge this characterization by defining CP as (p. 453) "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".

At the same time as the influential paper of Macmillan *et al.* appeared, Anderson *et al.* (1977b) [2] proposed a trainable neural-network model for associative memory (the so-called *brain-state-in-a-box*) jointly based on neurophysiological considerations [3] and linear systems theory, and applied it to categorical perception. Simulated labeling ('identification') and *ABX* discrimination curves were produced, and shown to be very much like those published in the experimental psychology literature. In the present paper, we refer to such use of neural-network models as *synthetic* CP. Unlike psychophysical studies with *real* subjects, which have maintained a constant high profile, synthetic CP (with some important exceptions, reviewed below) has remained largely unexplored [4]. This is in spite of the vast and continuing up-surge of interest in neural models of perception and cognition from ten 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 discover their operational principles; a point more recently made by Hanson and Burr (1990, p. 471) who state: "connectionist models can be used to explore systematically the complex interaction between learning and representation".

The purpose of this paper is to assess neural-network models of CP, with particular reference to their ability to simulate the behavior of real observers – including the relation between identification and discrimination. As is the case with any psychophysical model, the points in which synthetic CP agrees with observation can be taken as an existence proof that it is at least possible for real perceptual and cognitive systems to operate on the same principles as those embodied in the model. Conversely, the points in which real and synthetic behaviors differ may suggest worthwhile avenues for new and revealing experimentation. Hence, it is clear that synthetic CP is a worthwhile topic for study. In our work, we emphasize the importance of CP as an emergent property of a general learning system. Such an approach, unlike modeling based on curve-fitting, can suggest how the observed behaviors came to be as they are.

The remainder of this paper is structured as follows. In section 2, we outline the ways that CP has been characterized in the past, which reflect the historical development of the phenomenon and its psychophysical basis. Section 3 then reviews the various neural-network models which have been employed in the study of synthetic CP. These have mostly considered artificial or novel continua, whereas most experimental work with human subjects has considered speech, most often syllable-initial stop consonants. Thus, section 4 describes the application of two rather different neural systems to model the perception of such stimuli. Section 5 then considers the application of signal detection theory to synthetic CP, before concluding (section 6).

2 Characterization of CP

Categorical perception is usually defined or characterized with respect to some theoretical position. It is only natural, therefore, that views of CP have evolved in step with perceptual theories. However CP is defined, the relation between discrimination and identification remains key.

2.1 Early Characterizations from Speech CP

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

Liberman *et al.* (1957) investigated the perception of the plosive voicing contrast in syllable-initial stop consonants varying in place of articulation cued by second-formant transition while, shortly after, Liberman *et al.* (1958) studied the same contrast cued by first-formant 'cutback', simulating differing voice onset time (VOT). 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. An important finding in the latter case was that the phoneme 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 ms VOT through about 35 ms to approximately 42 ms. In contrast to the work with stop consonants, Fry *et al.* (1962) found the perception of vowels to be much ‘less categorical’.

At this time, and (in some quarters) for some years afterwards, CP was thought to reflect a mode of perception ‘special’ to speech (e.g. Liberman *et al.*, 1967; Studdert-Kennedy *et al.*, 1970) in which reference was made by the listener to articulation. It was supposed that an early and irreversible conversion took place of the continuous sensory representation to a discrete, symbolic code subserving both perception and production. Thus, perception of consonants is supposedly ‘more categorical’ than that of vowels just because the articulatory gestures producing the former are thought to be more discrete than the relatively continuous variations involved in production of the latter.

This ‘motor theory’ was partly predicated on the notion that Weber’s law was somehow fundamental, so that its violation necessitated special explanation. However, according to Macmillan *et al.* (1987, p. 32): “all … discrimination data require psychoacoustic explanation, whether they resemble Weber’s Law, display a peak, or are monotonic.” In spite of attempts to modify it suitably (e.g. Liberman and Mattingly, 1985; 1989; Liberman, 1996), the ‘speech is special’ hypothesis has been largely unable to withstand the weight of accumulating contrary evidence. See Schouten (1980), Rosen and Howell (1987) and Damper *et al.* (submitted) for extensive argumentation, review material and literature citations to this effect.

The 'classical' characterization of CP which emerges has been neatly summarized by Treisman *et al.* (1995, p. 335) as encompassing four features:

> "... a sharp category boundary, a corresponding discrimination peak, the predictability of discrimination function from identification, and resistance to contextual effects."

Subsequently in their paper (pp. 336–7), these authors critically assess this characterization, referring to: "... the unresolved difficulty that identification data usually predict a lower level of discrimination than is actually found" (e.g. Liberman *et al.*, 1957; Studdert-Kennedy *et al.*, 1970; Macmillan *et al.*, 1977; Pastore, 1987a) and the necessity to qualify "Studdert-Kennedy *et al.*'s claim that context effects *[and other sequential dependencies]* are weak or absent in categorical perception" (see also Healey and 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 and Darwin, 1978; Diehl *et al.*, 1978; Rosen, 1979; Repp and Liberman, 1987; Diehl and Kluender, 1987) which can shift the category boundary.

2.2 Signal Detection, Adaptation-Level and Criterion-Setting Theories

The pioneering Haskins work on phonetic categorization took place at a time when psychophysics was dominated by threshold models, and before the influence of signal detection theory (Green and Swets, 1966; Macmillan and Creelman, 1991) was fully felt. Broadly speaking, analyses based on signal detection theory (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 detailed way that discrimination is predicted from identification – with labeling playing a central role in both aspects of performance in the classical view. We deal with this latter point in some detail below, but make some brief remarks on the first point here.

***** FIGURE 1 ABOUT HERE *****

Figure 1(a) (after Massaro, 1987b[5]) shows the transformation of stimulus to response in an identification or discrimination task as a two-stage process: sensory operation followed by 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 current context is that classical notions of CP are ambiguous about *which* of the representations are categorical. That is, the information passed between sensory and decision processes (labeled X in the figure) could be categorical or continuous. In the latter case, this still allows for the response to be categorical, but this is *not* CP in Massaro's view since the categorization does not occur at the sensory/perceptual stage: he prefers the term *categorical partition*. Emphasizing this distinction, Massaro (1987a, p. 115) 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". (See also Hary and Massaro, 1982).

SDT avers that X represents a *continuous* decision variate. In the simplest single-interval, two-class case, X is unidimensional. The two classes of presentation give rise to two normal distributions of equal variance, one with mean 0 and the other with mean d'. However, as detailed by Macmillan *et al.* (1977), and reviewed below, the paradigms used in the study of CP have generally been more complex than this. Other important SDT analyses of CP include those of Elman (1979), Rosner (1984), Schouten and van Hessen (1992), van Hessen and Schouten (1992), Iverson and Kuhl (1995) and Takagi (1995) working with a variety of speech sounds.

Wilson (1987) interprets CP in terms of the adaptation-level (AL) theory of Helston (1964). This (quantitative) theory postulates that, as a result of experience, an organism is adapted to some 'neutral' level of stimulation which is set as:

$$\log A = p\log X + q\log B + r\log R$$

where A is the adaptation level, X is the geometric mean of the stimulus values presented for judgement, B is an 'anchor' stimulus value, R is a 'remote' or residual anchor value representing the long-term effects of previous exposure, and p, q and r are weighting coefficients that sum to unity. According to Wilson (p. 390), AL theory encompasses many important phenomena in psychophysics, including range-frequency effects. In the context of CP, the adaptation level is directly equated with the category boundary. Based on this, she shows that shifts in the boundary can be explained as a function of changes in adaptation level.

AL theory and SDT are similar in that they both posit an internal criterion against which category judgements are made. How then do they differ? Wilson nicely summarizes (p. 391) the relation between them as follows:

1. They share the proposition that CP is not to be contrasted with continuous perception. Rather, both stress the relation between identification and discrimination.

2. Both deny that categorical judgement is mediated by labeling which is "a rather special claim in any case, as it rests on a particular experimental paradigm".

3. Exponents of both theories agree that "finding discontinuities in sensitivity ... is no more than a curiosity until the underlying mechanism can be spelled out."

A similar idea to AL theory's notion of shifts in the internal criterion as experience changes is embodied in the recent work of Treisman *et al.* (1995). Starting from Elman's (1979) SDT analysis, and which "suggested that consonant adaptation effects might be due to criterion shifts", they propose a so-called criterion-setting theory (CST). 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 in CP. When applied to *ABX* discrimination, the theory "is shown to fit the data from the literature" (p. 334), in that a peak occurs at the category boundary. Further, the absolute value of the observed discrimination performance is close to that predicted by CST. This is not the case with the standard, or classical Haskins model, which predicts (from labeling) a lower performance than is actually observed, as we now discuss.

2.3 Prediction of Discrimination from Identification

In the classical Haskins view, discrimination in an *ABX* task (as traditionally employed in CP studies) is a process of covert labeling. That is, first A is labeled 'covertly'[6], 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. Based on this view, *ABX* discrimination is predictable from identification[7]. For judgements involving just two categories, where the prior probability of each is equal, the proportion correct in discrimination is predicted as (Liberman *et al.*, 1957; Macmillan *et al.*, 1977):

$$P(C) = 0.5\left[1 + (p_A - p_B)^2\right] \tag{1}$$

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 0.5. It is well-known that this model leads to a predicted discrimination which (as stated above) is almost invariably lower than the observed discrimination. CP theorists have usually somewhat rationalized this discrepancy by emphasizing the correlation between the predicted and observed curves, i.e. their similar, non-monotonic shape and the fact that they peak at approximately the same (boundary) point.

Massaro (1987a, p. 91) writes: "For some reason, the discrepancy has never been a deterrent for advocates of categorical perception nor a central result for any alternative view". (See also Massaro, 1987b, p. 256.) 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 co-exist (Figure 1(b)). If the subject fails to label A and B differently via the categorical route then, instead of guessing, the continuous (but decaying) representations of A and B are consulted. According to Macmillan *et al.* (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 make the point that – by introducing an extra process (and an attendant free parameter) – the theory must necessarily provide a better fit between discrimination and identification. Hence, the improved fit can in no sense be taken as support for the hypothesis of dual processes [8]. Further, they criticize dual-route theory for "its embarrassing lack of parsimony" (p. 467) in that everything that can be done via the discrete route, and more, can also be achieved via the continuous route.

Macmillan *et al.* point out that the Haskins model is tacitly based on "low threshold" assumptions (see also Pastore, 1987b), and argue 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 signal detection theory, in terms of the equivalence of discrimination and identification. The essential defining characteristic of CP is then considered to be equivalence of identification d', found using the approach proposed by Braida and Durlach (1972) for auditory intensity perception, and discrimination d'. However, the proportion correct measure, $P(C)$, employed in the classical *ABX* discrimination task extensively employed in CP studies does not, of course, yield an unbiased estimate of sensitivity. Thus, Macmillan et al 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, in order to extract unbiased sensitivity measures in this latter case. The *ABX* task was analyzed (pp. 458–9) as a *2IFC* sub-task (to determine if the standards are in the order *AB* or *BA*), followed by a yes-no sub-task. This is described (p. 462) as "a continuous (SDT) model for categorical perception" [9].

A final contribution of this important publication was to introduce the *response-constraint hypothesis*, according to which sensitivity in identification tasks can be seriously under-estimated if there are too few response categories – so that most stimuli can be identified without error. This leads to the possibility of experimental test of a response-constraint versus the dual (auditory plus phonetic) processing hypothesis, by varying the number of response categories available to subjects. In short, CP should ideally be tested using multi-response identification tasks and, preferably, a two-interval forced choice discrimination task from which 'true' d' can be obtained in unbiased form. We do not pursue this here – primarily because this suggested approach has not in fact superseded the more classical paradigm and we wish to consider synthetic CP within the context of existing experimental results.

3 Neural Models of CP: A Review

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

3.1 The 'Brain-State-in-a-Box'

Early neural models of categorical perception were essentially based on associative memory networks, since this was one of the few kinds of network that was attracting any kind of interest in the aforementioned "dark ages" (see note 4) before the discovery of the error back-propagation algorithm (Rumelhart *et al.*, 1986). (See Kohonen, 1977, and Hinton and Anderson, 1981, for extensive contemporary reviews of parallel models of associative memory [10] and Anderson, 1995, for a more recent introductory treatment.). This is quite a natural model for CP in many ways. An associative memory network 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 non-prototypical stimulus is replaced in memory by (and so is indistinguishable from) its prototype.

We will take Anderson *et al.*'s (1977b) paper as the starting point for our review of neural models of CP. We note, however, that this selection is possibly contentious. Grossberg, for instance, also has a legitimate claim to be the originator of this line of research with his very early papers on neural models with psychological implications (Grossberg, 1968a; b; 1969)[11]. With due respect to Grossberg, our preference for the Anderson *et al.* model is based on its greater simplicity and perspicacity, and its more direct and obvious use to model human psychophysical data, but partly also because Anderson's work on associative memory dates back at least as far (see Anderson, 1968).

In their 1977 paper, Anderson, Silverstein, Ritz and Jones consider networks of neurons[12] which (p. 416) "are simple analog integrators of their inputs". 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 usefully represented by the pattern of activation across a group of cells;
2. different memory traces make use of the same synapses;
3. synapses associate two patterns by incrementing synaptic weights in proportion to the product of pre- and post-synaptic activities.

The form of learning implied in 3 is in effect correlational, and has been called 'Hebbian' by many workers. As the neurons have linear activation functions, a form of linear correlation is performed, making the network amenable to analysis using 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 network is created with N input units and M output units. In accordance with 2 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:

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

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:

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

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 is, however, 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 network will produce noise as well as the correct output but will still (p. 417) be "quite usable".

To convert this linear pattern-association network into a model of CP, Anderson *et al.* 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 (non-orthogonal) inputs, it is shown (p. 424) that (provided their *average* is zero) the inputs are a linear combination of the eigenvectors of the feedback matrix $\mathbf{A}$, and all eigenvalues are positive.

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 units are eventually driven into saturation (either positive or negative in sense) and the network has stable states corresponding to some (possibly all) of the 'corners' of a hypercube in its N-dimensional state space [13]. For this reason, the model was called 'brain-state-in-a-box' (BSB). 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 (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 categorical perception, the inputs $\mathbf{f}_i$ can be associated with themselves during training, i.e 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, of course, be similarly labeled according to the input patterns that they attract [14]. 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 which is a canonical or prototypical form corresponding to the eigenvector of the input with the largest eigenvalue. Anderson *et al.* (pp. 430–433) present a simulation of categorical perception in which their neural model performed two-class identification and *ABX* discrimination tasks. The two 'prototypes' (eigenvectors) were the 8-dimensional orthogonal vectors of length-2 in the directions $(1, 1, 1, 1, -1, -1, -1, -1)$ and $(1, 1, -1, -1, 1, 1, -1, -1)$ respectively (Figure 2(a)). These were used to set the weights. Inputs to the model then consisted of 100 repetitions of 16 length-1 vectors equally spaced between the prototype eigenvectors, with added random noise according to one of four conditions: the standard deviation (SD) of the noise was either 0.0, 0.1, 0.2 or 0.4.

We have repeated Anderson *et al.*'s 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 1000 presentations under three different noise conditions. During testing, the activation of each neuron was computed as:

$$act_i(t) = \alpha \left(extinput_i \right) + \beta \left(intinput_i \right) \tag{3}$$

where $extinput_i$ and $intinput_i$ are, as their names clearly suggest, the external input to unit i and the internal, fedback input to the same unit. A decay term can be incorporated into the model, which tends to restore activation to a resting level of zero:

$$\Delta act_i = \alpha \left(extinput_i \right) + \beta \left(intinput_i \right) - (decay) act_i$$

Throughout this work, *decay* was set to 1 so that the activation is given simply by equation (3).

For the repetition of Anderson *et al.*'s simulation, the external scale factor α and internal scale factor β were both set at 0.1. 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 did Anderson *et al.* [15]. Thus, our noise conditions were $SD = 0.0$, 0.4 and 0.7.

***** FIGURE 2 ABOUT HERE *****

In all noise-free cases, the system converged to one of its two stable, saturating states for all 1000 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 $SD = 0.7$. Figure 2(b) shows the identification results obtained by noting the proportion of inputs which converge to the saturating state corresponding to endpoint 0. For the no-noise condition, categorization is perfect with the class boundary at the mid-point between the prototypes. For the noise conditions $SD = 0.4$ and $SD = 0.7$, the labeling curves are very reasonable approximations to those seen in the classical CP literature. Overall, this replicates the essential findings of Anderson *et al.*

Consider next the *ABX* discrimination task. Anderson *et al.* considered two inputs to the network to be discriminable if they converged to different stable states. If they converged to the same stable state, a guess was made with probability 0.5, in accordance with equation (1). This means that discrimination by the net is effectively a direct implementation of the Haskins model; hence, we expect *real* discrimination performance to be underestimated as it is with the Haskins model. Indeed, they observed a distinct peak at mid-range 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 and 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" in the case that the inputs are not orthogonal, Anderson (1977, pp. 78–83) presents an example in which the BSB model is used to categorize vowel data (see also Anderson, Silverstein and Ritz, 1977a). 12 Dutch vowels were represented by 8-dimensional vectors, where each element measures the energy within a certain frequency band of an 'average', steady-state vowel. It is highly unlikely that these inputs are mutually orthogonal, yet (p. 81) "when learning ceased, each vowel was assigned to a different corner". Indeed, as mentioned earlier, non-orthogonality can act as 'noise', so preventing (unrealistic) perfect categorization.

Anderson *et al.* (1977b, p. 433) conjecture that positive feedback, saturation and synaptic learning are "responsible for the interesting *[categorization]* effects in our simulations". With the benefit of hindsight, however, we now know (see extensive review material below) that synthetic CP can be obtained in a variety of neural models – even those lacking positive feedback and saturation. In this regard, the comments of Grossberg (1986, pp. 192–194) concerning saturation in the BSB model are apposite. He charged Anderson *et al.* with introducing a homunculus, as a result of their "desire to preserve the framework of linear systems theory", going on to write:

> "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."

In our view, this charge is without foundation. A homunculus is some unjustified, implicit mechanism which is, in the worst case, comparable in sophistication and complexity to the phenomenon to be 'explained'. By contrast, Anderson *et al.* postulate an explicit mechanism – that of firing-rate saturation – which is both simple and plausible in that something like it is a ubiquitous feature of all neural systems [16]. To be sure, the discontinuity of the linear-with-saturation activation function is mathematically unsatisfactory, but essentially similar behavior is (as detailed below) observed in neural models with activation functions having a more gradual transition into saturation.

3.2 The TRACE Model

In 1986, McClelland and Elman produced a detailed, connectionist model of speech perception, featuring localist representations and extensive top-down processing in addition to the more usual bottom-up flow of information. 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, while those within a level that are contradictory have inhibitory connections – that is, processing is *competitive*. Strictly, 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 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 [17] (and, indeed, about whether this an appropriate representation or not.) Top-down effects are manifest through the lexical status (or otherwise) of words affecting phoneme 'perception' and thereby feature 'perception' also. Although TRACE has been used to simulate a variety of effects in speech perception, we concentrate here on its use to model CP.

***** FIGURE 3 ABOUT HERE *****

An 11-step /g/–/k/ continuum was formed by interpolating the distinctive feature values. Features used were voice onset time (VOT) and the onset frequency of the first formant. 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 influence, and all phoneme units other than /g/ and /k/ were also removed. Figure 3(a) shows the initial activations (at time step $t = 1$) at these two units as a function of stimulus number. As can be seen, there is a clear trend for the excitation (which is initially entirely bottom-up) to favor /g/ at low stimulus number but /k/ at high stimulus number. The two curves cross at stimulus number 6, indicating that this condition is maximally ambiguous (i.e. this is the phoneme boundary). However, the variation is essentially continuous rather than categorical as shown by the relative 'shallowness' of the curves. By contrast, after 60 time steps, the two representations are as shown in Figure 3(b). As a result of mutual inhibition between the /g/ and /k/ units, and possibly of top-down influence of phoneme units on featural units also, a much steeper, 'more categorical' response is 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, p. 239) states:

> "Categorical perception can ... be anticipated whenever adaptive filtering interacts with sharply competitive tuning, not just in speech recognition experiments (Hary and Massaro, 1982; Pastore, 1981; Studdert-Kennedy, 1980)."

McClelland and Elman 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 3(c), 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. Accordingly, Quinlan (1991, p. 151) 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."

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.

Finally in this subsection, we consider the relation between discrimination and identification in TRACE. McClelland and Elman (p. 47) point out the actual discrimination in *real* CP is better than predicted from identification and write that TRACE also "produces this kind of approximate categorical perception". The mechanism by which this happens is 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, while 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.) However, this explanation relies upon interrogating the model *before* the interactive process has run to completion: how do we know precisely when to do this? The authors remark on the practical difficulty of distinguishing between this feedback explanation and a dual-process explanation.

3.3 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 networks 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 categorical perception of both speech (dealt with in the next section) and artificial stimuli (dealt with here).

Many workers (Bourland and Kamp, 1988; Elman and Zipser, 1988; Baldi and Hornik, 1989; Hanson and Burr, 1990) have observed that feedforward 'auto-associative' networks [18] with hidden units effectively perform a principal component analysis of their inputs. Harnad *et al.* (1991) exploited auto-association training to produce a "precategorization" discrimination function. This was then re-examined after categorization training to see if it had 'warped'. That is, a CP effect was defined as a decrease in within-category inter-stimulus distances and/or an increase in between-category inter-stimulus distances relative to the auto-association-alone baseline. The stimuli studied were artificial – namely, different representations of the 'length' of 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 8 different input 'lines' were represented in 6 different ways, to study the effect of the "iconicity" of the input coding, i.e how analog, non-arbitrary, or structure-preserving it was in relation to what it represented. After auto-association training (using 8 output units), the trained weights between hidden layer and output layer were re-loaded, the input to hidden layer weights 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 were observed for all input representations: the more "iconic" the representation, the stronger the effect. The categorization task was very difficult to learn with only 2 hidden units. Provided there were at least 3 hidden units, however, synthetic CP behavior was not at all sensitive to their number. This is taken to indicate that CP is not merely a by-product of information compression by the hidden layer. Nor was CP a result of over-learning to extreme values, since 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 if CP was an artifact of re-using the weights for the precategorization discrimination (auto-association) for the auto-association-plus-categorization. Performance was averaged over several precategorization nets, and compared to 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 "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". In this and subsequent work (Harnad *et al.*, 1995), the time-course of the training was examined and three factors playing an important part in generating synthetic CP were identified:

1. maximal interstimulus separation induced during auto-association learning – according to which the hidden-unit representations of each (initially random) stimulus move 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 1 in a way which promotes within-category compression and between-category separation;

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 closer the initial representation is to satisfying the partition implied by 1–3 above, the more pronounced will be the network's categorization behavior.

From this work, it is apparent that a feedforward neural network, trained on back-propagation, is able (in spite of its obvious dissimilarities) to replicate the essential features of classical CP in much the same way as the BSB model of Anderson *et al.* (1977b). There are, however, noteworthy differences. The most important one is that the back-propagation net is, in effect, explicitly *told* where to place the category boundary during categorization training, rather than generalizing the correct boundary from training on the stimulus-continuum endpoints, so that its CP behavior is arguably not "emergent" at all. Also (because the feedforward network has no dynamic behavior resulting from feedback), reaction times cannot be quite so easily predicted as Anderson *et al.* do (but see section 4.3).

3.4 Competitive Learning and Category Detecting Neurons

Goldstone *et al.* (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') between two randomly-selected bezier curves. The dimension is novel in that subjects are highly unlikely ever to have seen precisely these morphed shapes before. The major interest in the context of this paper is that these authors also present a neural-network model (a form of radial-basis function network) which qualitatively replicates the behavioral results.

The model has a layer of hidden neurons which become specialized for particular stimulus regions: so 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 adjustment of the input-to-hidden (or 'position') weights. Simultaneously, associations between hidden/detector neurons and output (so-called category) units are learned by gradient descent. As well as 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 to a detector that produces an improper categorization.

Goldstone *et al.* 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" (Kruschke, 1992, 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 [19]. (Note that, like the back-propagation model of Harnad *et al.*, the category boundary is explicitly placed, so that the categorical behavior is not truly emergent.) The labeling probability showed a characteristic warping, with its 50% point 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.

The authors contrast their work with that of Anderson *et al.* (1977b) and Harnad *et al.* (1995). In these other approaches, they say (p. 248), "each category has its own attractor" so that CP "occurs because inputs that are very close but fall into different categories will be driven to highly separated attractors". In their network, 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 fashion in CP research which holds that the paradigm "has overemphasized the importance of the phonetic boundary between categories" (Repp, 1984, p. 320) and seeks instead to explore the internal structure of the categories in terms of anchors and/or prototypes (e.g. Macmillan, 1987; Kuhl, 1991; Volaitis and Miller, 1992; Miller, 1994; Iverson and Kuhl, 1995).

4 Categorization of Stop Consonants by Neural Networks

From the foregoing review, it is apparent that neural models of CP have no real problem replicating the classical observations of a sharp labeling function and a peaked discrimination function. While 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 networks and approaches to support the claim that they reflect the emergent behavior of any reasonably powerful learning system. With the exception of the vowel categorization work using the BSB model (Anderson, 1977; Anderson *et al.*, 1977a), however, the neural models of synthetic CP reviewed thus far have all taken their inputs from artificial or novel dimensions. Yet the vast majority of real CP studies have used speech stimuli – most often stop consonants (or, more correctly, simplified analogs of such sounds). Hence, our goal for this section is to consider the categorization of stop consonants by a variety of neural-network models. As mentioned in section 2.1, 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 if neural models of CP reproduce this effect as emergent behavior.

4.1 Stimuli and Pre-Processing

The stimuli used in this section were synthesized consonant-vowel syllables supplied by Haskins Laboratories. They are digitally-sampled versions (sampling rate 10 kHz) of those developed by Abramson and Lisker (1970) and employed extensively in previous studies of speech CP. They consist of three sets in which VOT varies in 10 ms steps from 0–80 ms, simulating a series of English, pre-stressed, bilabial (/ba–pa/), alveolar (/da–ta/) and velar (/ga–ka/) syllables.

The stimuli were pre-processed for presentation to the various network models using a computational model of the peripheral auditory system (Pont and Damper, 1991)[20]. Full details of the pre-processing are given elsewhere (Damper *et al.*, 1990; 1996; submitted): only a brief and somewhat simplified description follows. The output of the auditory model is a '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. Because of the so-called 'tonotopic' (frequency-place) organization of auditory nerve fibers, and because the simulated fibers are systematically spaced 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 ms to 95 ms in 10 ms steps in the time dimension and from 1 to 128 in steps of 8 in the frequency (fiber index) dimension: thus, the networks have a maximum of 192 inputs. These time limits were chosen to exclude most (but not all) of the pre-stimulus spontaneous activity and the region where responses were expected to be entirely characteristic of the vowel.

Since the auditory pre-processor 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 pre-processor repeatedly with the same input. In this work, 50 repetitions were used for each of the three (bilabial, velar, alveolar) series to produce neural spectrograms for training and testing the networks.

4.2 Brain-State-in-a-Box Model

The input data were first reduced to bipolar, (approximately) zero-mean 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 were run using 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 ms VOT, and canonical unvoiced or 80 ms 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 a N-input, N-unit BSB network: this is a form of orthogonalization [21]. 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 [22]. The network was then trained on the 0 ms and 80 ms endpoints, and generalization was tested on the full range of stimuli including the (unseen) intermediate (10–70 ms) stimuli.

Because of the relative 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 $C = \pm 20000$). 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 ms (voiced) training patterns were always assigned different corners from the 80 ms (unvoiced) patterns. During generalization testing, no rubbish states were encountered – convergence was always to a labeled attractor. Further, the activation vector (after convergence) for the 0 ms 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 ms (unvoiced) endpoint stimuli. (This would, of course, be a problem if the task of the network was to identify the place of articulation rather than the presence/absence of voicing.)

***** FIGURE 4 ABOUT HERE *****

Figure 4(a) shows the obtained labeling function (averaged over 50 presentations), while Figure 4(b) shows the obtained 1-step discrimination function (averaged over 1000 presentations). The results are clear and unequivocal. Classical categorization is observed with a steep labeling curve and an *ABX* discrimination peak at the category boundary. While the labeling curve is rather too steep and the actual boundary values obtained are slightly low (by about 5 or 10 ms), the shift with place of articulation is qualitatively correct. It is also worth noting that with these more realistic input patterns, there is no need to add noise as there was in the case of the ‘artificial’ (vectors) input.

***** FIGURE 5 ABOUT HERE *****

Figure 5 shows the obtained 1-step discrimination function for the alveolar series, and that predicted on the basis of equation (1). They are essentially identical, differing only in the fact that the obtained function is contaminated by the sampling statistics of the guessing process.

4.3 Back-propagation Network

In light of the foregoing review (Section 3), 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, the network is trained on the 0 ms and 80 ms endpoints and generalization is then tested using the full range of VOT stimuli.

2. Using the auto-association paradigm of Harnad *et al.* (1991; 1995), hidden-unit representations resulting from pre- and post-categorization training are compared.

In this work, we have adopted approach 1, mostly because boundary placement is more obviously an emergent property in this case [23]. We leave approach 2 for future work.

Initially, one MLP was constructed for each of the three (bilabial, alveolar, 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 [24]: two were used in the following. Suitable training parameters (arrived at by trial and error) for $n = 2$ were:

- learning rate, $\eta = 0.005$;
- momentum = 0.9;
- weight range = 0.05;
- $\epsilon = 0.25$.

The ϵ error criterion was arrived at by allowing an average error of 0.05 (or 0.0025 when squared) for each of the 100 training patterns.

***** TABLE 1 ABOUT HERE *****

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 0.25 error criterion very easily: typically in about 50 epochs. The strong tendency, especially for those cases where the criterion was reached quickly, was to encode the 0 ms endpoint with hidden unit activations of $h_1 h_2 = 01$ and the 80 ms endpoint with $h_1 h_2 = 10$. (Of course, h_1 and h_2 were never *exactly* 0 or 1 but, more typically, something like 0.05 or 0.95.) On only one exceptional occasion (when training required 230 epochs) was a hidden-unit coding arrived at for which h_1 and h_2 for the different endpoints were not *both* different. Similar results were obtained for the two other (alveolar, velar) nets, except that the alveolar net was rather more inclined to discover the $h_1 h_2 = 00/11$ coding. Weight sets selected for subsequent testing were those obtained in the minimum number of training iterations (these all embodied the $h_1 h_2 = 01/10$ coding).

Figure 6(a) shows the labeling function obtained by averaging output activations [25] over the 50 stimulus presentations at each VOT value, for the three nets. Again, classical CP was observed, with a steep labeling function and separation of the three curves according to place of articulation. The labeling curve for the bilabial series is less good than for the alveolar and velar stimuli, with the average activation being rather too low for the 20 ms VOT and somewhat too high for VOTs greater than 30 ms [26]. The boundary movement with place of articulation is an emergent property of the nets – or, perhaps more strictly, of the way they learn to process their (different) inputs. There is no sense in which the nets are explicitly trained to separate the boundaries in this way.

***** FIGURE 6 ABOUT HERE *****

Consider now the discrimination functions for the MLPs. Unlike the Anderson *et al.* simulation which produces a discrete code, the MLPs produce a continuous value in the range (0,1) because of the sigmoidal activation function. This means 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* sub-task to determine the order of the standards, *AB* or *BA*, followed by a yes-no sub-task. The *A* and *B* patterns were selected at random from adjacent classes, i.e. from the sets of 50 responses at VOT values differing by 10 ms. The *X* pattern was chosen at random, with equal probability of 0.5, from one of these 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 sub-tasks into one. Let the absolute difference in activation between the *X* and *A* inputs be $|X - A|$; similarly $|X - B|$. The classification rule is then:

$$X \text{ is } \begin{cases} A & \text{if } |X - A| < |X - B| \\ B & \text{otherwise} \end{cases} \tag{4}$$

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

We found, however, $|X - A|$ and $|X - B|$ were occasionally identical in our simulations. In terms of the Macmillan *et al.* model, this means that the *2IFC* sub-task would have yielded *AA* or *BB* outcomes which are inappropriate for the yes-no sub-task. To avoid making the simulation overly sensitive to round-off errors, we therefore introduced a 'guessing threshold', g. According to this, *X* was only classified by rule (4) above if:

$$\big||X - A| - |X - B|\big| < g$$

If this inequality was not satisfied, then the classification of *X* was guessed with an equal probability of 0.5 for each class. Results were not especially sensitive to the actual value of g. (It is important to realize that g is in no sense an additional 'free' parameter which improves the fit between the simulation and real data.)

Figure 6(b) shows the discrimination functions obtained from such a simulated *ABX* experiment. There were 500 $\langle ABA \rangle$ presentations and 500 $\langle ABB \rangle$ presentations, 1000 in total. Again taking advantage of a computational short-cut, there were no presentations of the BA standard, on the grounds that the simulation had perfect 'memory' so that symmetry is assured and this condition is practically indistinguishable from the AB standard. The guessing threshold, g, was 0.001. There are clear peaks at the phoneme boundary, and the movement of these peaks with the place of articulation is qualitatively correct. Paralleling the less-steep (and more psychophysically reasonable) labeling functions, the discrimination peaks are not as sharp as for the BSB model. They are closer to those typically obtained from real listeners.

***** FIGURE 7 ABOUT HERE *****

Figure 7 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 whereby observed discrimination in psychophysical tests exceeds that predicted from equation (1). This shows the importance of continuous representation 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 6 endpoint stimuli, it will not generalize so as to put the boundary in the correct location for each of the three series of inputs. Work is continuing to see if this problem can be overcome using more hidden units and/or a different output coding.

***** FIGURE 8 ABOUT HERE *****

Figure 8 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 – see note 26). Indeed, the standard deviation could be taken as a credible predictor of reaction time in human psychophysical experiments.

5 Synthetic CP 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 using the techniques that 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 signal detection in particular are relevant to the evaluation of neural-network 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 network.) 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 while false alarms are 'yes' responses to signal-plus-noise presentations. In the *ABX* discrimination paradigm, it is arbitrary whether $\langle ABA \rangle$ and $\langle BAB \rangle$ are taken to correspond to signal presentations, while $\langle ABB \rangle$ and $\langle BAA \rangle$ are taken to correspond to signal-plus-noise, or *vice versa*. Following Macmillan *et al.* (1977) and Macmillan and Creelman (1991), we adopt the former convention, as in Table 2 where 'Response 1' means that the subject nominated the first interval as containing the standard corresponding to *X*. ('Response 2', of course, 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 previously stated, only the stimulus-response matrix in the top half of the table was collected.

***** TABLE 2 ABOUT HERE *****

Macmillan *et al.* (1977, p. 459) 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" (their equation (3)):

$$\begin{aligned} H &= P(\text{'Response 1'}|\langle ABA\rangle) = P(\text{'Response 1'}|\langle BAB\rangle) \\ &= 1 - FA \end{aligned}$$

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 *et al.* (1978). In the terms of Macmillan *et al.* (1977), CP requires that this true d' for discrimination is 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. Second, it is inherently unsuitable for the analysis since its labeling function includes many 1 and 0 points, which are not amenable to transformation to z-scores as they yield values of $\pm\infty$.

With the simulation data of the previous section, our first step was to check the unbiased assumption, according to which $H + FA = 1$. Taking the representative case of the alveolar series, the mean value of $H + FA$ was 0.9895 with standard deviation 0.26483. A t-test ($t = 0.105$ with 6 degrees of freedom) reveals that this is not significantly different from 1 at even the 90% level. Comparable results were found for the other VOT series, indicating that the unbiased assumption is tenable and implying that it is immaterial whether we use $2z(H)$ or $z(H) - z(FA)$ as the measure of d'_s. To check this, a paired t-test of $2z(H)$ and $z(H) - z(FA)$ was carried out to test the null hypothesis of equality. For the alveolar series, this gave $t = 0.0007$ (with $\nu = 7$ degrees of freedom, two-sided test) giving no basis whatsoever to reject the null hypothesis. Again, similar findings obtained for the other two VOT series. All subsequent work uses $2z(H)$ as the measure of d'_s.

***** TABLE 3 ABOUT HERE *****

***** FIGURE 9 ABOUT HERE *****

Table 3 shows the (one-step) discrimination d', found using equation (5) and Table 3 of Kaplan *et al.*, and the identification d' obtained from the z-transformed identification proportions for the representative case of the alveolar series. In Figure 9, discrimination d' is plotted against identification d'. According to the Macmillan *et al.* characterization of CP, these two measures should be equivalent. That is, the points of Figure 9 should be distributed about the unit-slope straight line. Clearly, discrimination d' exceeds identification d' by some margin. On the basis of a paired t-test ($t = 3.0790$ with $\nu = 7$ degrees of freedom, two-sided test), these two sensitivity measures are significantly different from equivalent at the 2% level. Similar findings hold for the bilabial and velar series. However, the two measures are highly correlated. Regression analysis (alveolar series) yields $r = 0.756$ for the case when (on the basis of the assumed equality) the best-fit line is forced to pass through the origin and $r = 0.991$ when it is not. The former is significant at the 2% level.

Hence, we conclude that the back-propagation neural model does not exhibit categorical perception in the sense of equivalence of discrimination and identification. Rather, like human observers (Macmillan, 1987, p. 59, discrimination performance is correlated with but higher than identification performance

6 Conclusions and Discussion

The use of trainable neural-network models is an under-explored yet insightful way of studying the important phenomenon of categorical perception. In this paper, we have shown that a variety of neural models is capable of replicating classical CP, whereby the midpoint of the steep labeling function and sharp peak of the discrimination function are coincident at the category boundary. Given the ubiquitous way that CP arises in network behavior, 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 important special case of speech CP, we have also demonstrated the shift of phoneme boundary with place of articulation for the voicing continuum using two different networks – the historically-important brain-state-in-a-box model of Anderson *et al.* (1977) and a more recent multilayer perceptron ('back-propagation') model. The most convincing demonstration of synthetic CP to date is by the perceptron model. The categorization behavior is an emergent property of the network 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 network also replicates the frequently-documented effect whereby observed discrimination performance exceeds that predicted from labeling on the assumption that only discrete information is available about the labels (the so-called Haskins model). It does so by retaining continuous information after the stage of sensory processing, i.e. the categorization occurs at the later decision stage.

Early classical studies of CP considered the effect in terms of low-threshold theories, and used proportion correct, $P(C)$, as a sensitivity measure in *ABX* discrimination tasks. This measure is, of course, not free of response bias on the part of the observer. In 1977, Macmillan *et al.* contributed an important advance by using the more modern methods of signal detection theory, which clearly separate measures of sensitivity (d') from response bias (β). They suggested that CP should be characterized by the equivalence of identification and discrimination sensitivity, both measured with 'true' d'. Our back-propagation simulations fail to satisfy this definition in that identification d' is statistically different from discrimination d', although highly correlated with it. This parallels the usual finding with real, human observers.

Since 1977, Macmillan has retreated somewhat from the notion of equivalence of discrimination and identification as the defining characteristic of CP (see also Macmillan *et al.* 1988). Thus, he writes (Macmillan, 1987, p. 78): "It is clear that few if any dimensions have this property". Nonetheless: "Relations between tasks *can* provide useful information about the manner in which stimuli are processed, however such processing is named". In this work, we have shown that the relation between identification and discrimination performance for a simple neural-network simulation closely parallels that seen for real subjects. As 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.

There is (at least) one way in which our simulations of 'synthetic' CP are too simple: 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 neural-network models.) That is, in the case of VOT perception, the reduced (192-value) 'neural spectrogram' is available as a static input to the back-propagation network. Apart from an implicit "time as space" representation, there is no explicit representation of relational time. 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. Obviously, real listeners hear a sequence of speech sounds, 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 Elman (1988) – so as to implement an imperfect memory buffer.

Acknowledgements

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.

Let's ask some kind people to look the paper over before we submit it . . .

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.

Notes

[1]Indeed, the title of the present paper is closely and deliberately modeled on that of Macmillan and his colleagues.

[2]The two papers actually appeared consecutively in the same issue of *Psychological Review*.

[3]Rather uncharitably, Remez (1987, p. 219) refers to the "neurophysiological speculations" of Anderson *et al.* – as if most of psychology was not speculation!

[4]Indeed, Dreyfus and Dreyfus (1988, footnote 8) credit James A. Anderson – along with Stephen Grossberg and Tuevo Kohonen – with keeping neural-network modeling in artificial intelligence and cognitive science alive "during the dark ages" in which it was considered by most workers in the field to been have superseded by the physical-symbol system hypothesis of intelligence.

[5]This figure is redrawn from Fig. 8.5, p. 262, of Massaro (1987b) but originates from Massaro (1975).

[6]'Covert' in the sense that the subject is not required to report this judgment to the investigator as in 'overt' identification.

[7]Indeed, one of the criticisms of this paradigm (e.g. Pisoni and Lazarus, 1974; Massaro and Oden, 1980) is that it promotes identification (labeling) behavior, so arguably promoting categorization behavior also.

[8]The theory does, however, have other strengths. It can explain the effect that memory requirements of the experimental procedure have on categorical perception on the basis that the two processes have different memory-decay properties. See Schouten and van Hessen (1992) and Damper *et al.* (submitted).

[9]This view – of the importance of continuous information to categorical perception – is gaining ground over the classical characterization of CP, c.f. Wilson's (1987) contention that CP is not to be contrasted with continuous perception. More recently, Treisman *et al.* (1995, p. 334) state that "CP resembles standard psychophysical judgements" while Takagi (1995, p. 569) writes: "In fact, the signal detection model is compatible with both categorical and continuous patterns of identification/discrimination data."

[10]Many readers will be aware of the dramatic effect that Hopfield's 1982 paper in the *Proceedings of the National Academy of Sciences, USA* had on the field of neural computing. It is worth asking why this was, considering that Hopfield's work had many similarities with that of Anderson, for instance, and – to that extent – was not revolutionary. Undoubtedly, one reason was Hopfield's reputation as a researcher outside of neural computing. Another was his introduction of the analogy between network dynamics in terms of 'energy' and the dynamics of general non-linear systems in terms of a Lyapunov function, which gave a new and insightful way to visualize network operation.

[11]See Grossberg (1987) and Carpenter *et al.* (1991a; b) for more recent developments.

[12]We use the terms *neuron* and *unit* interchangeably.

[13]Not all corners are necessarily stable.

[14]There is, of course, no guarantee that all corners will be so labeled. Corners which remain unlabeled correspond to 'rubbish states' in the jargon of associative networks.

[15]We believe this is because our use of 1000 test patterns in place of Anderson *et al.*'s 100 makes our results less affected by small-sample effects.

[16]In the words of Lloyd (1989, p. 205): "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". Anderson *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 Anderson and Silverstein's (1978) reply thereto, for additional discussion of the status of the non-linearity in the Anderson *et al.* BSB model, as well as Bégin and Proulx (1996) for more recent commentary.

[17] In their 1986 *Cognitive Psychology* paper, McClelland and Elman describe TRACE II which, they say (p. 13), is "designed to account primarily for lexical influences in phoneme perception" using "mock speech" as input. 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. Elman and McClelland imply (p. 364) that the stage of conversion from real speech input to "feature parameter values" is without difficulty, and was excluded "simply to speed up the simulation". We do not believe that many informed commentators would be inclined to agree with this sentiment.

[18]In this context, 'auto-associative' means that a multilayer feedforward network with hidden units is trained to reproduce its inputs as outputs. This use of the term should be carefully distinguished from the use to describe 'auto-associative' feedback networks (like the BSB model and the Hopfield network) in which there are no hidden units and the 'input' and 'output' units are physically identical.

[19]Although Luce's choice rule is apparently used in the Goldstone *et al.* model, it seems that the k parameter which was treated by McClelland and Elman as free in the TRACE model and adjusted to give CP, is here treated as fixed (at unity).

[20]The use of such sophisticated pre-processing obviously requires some justification. First, there has long been a view in the speech research literature that CP reflects some kind of "restructuring of information" (Kuhl and Miller, 1978, p. 906) by the auditory system in the form of processing non-linearities – see Damper *et al.* (submitted) for extensive discussion. Accordingly, we wished to find correlates of CP in the neural activity of

the auditory system, following Sinex and McDonald (1988, p. 1817) 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." Also, as a step towards understanding the acoustic-auditory restructuring of information, we wished to discover the important acoustic features which distinguish initial stops. In the words of Nossair and Zahorian (1991, p. 2990), 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". See Damper (1997) for the latest work in this vein.

[21]We believe that, ideally, this kind of pre-analysis is best avoided. The neural model *ought* to be powerful enough in its own right to discover the important inputs.

[22]These 20 'most important' time-frequency cells are, not surprisingly, located around the low-frequency region (corresponding to 118 to 740 Hz) just after acoustic stimulus onset where voicing activity varies maximally as VOT varies. The precise location of this region shifts in the three nets (bilabial, alveolar, velar) in the same way as does the boundary point. See Damper *et al.* (submitted) for detailed discussion of the importance of this region of the neural spectrogram.

[23]. . . but also because for VOT perception we have psychophysical data against which to assess our simulation. This was not the case for Harnad *et al.*'s artificial data, which accordingly required some other reference for comparison.

[24]In fact, we do not need hidden units at all. Damper *et al.* (submitted) show that synthetic CP of the VOT continuum is exhibited by single-layer perceptrons, and exploit this fact in identifying the neural correlates of CP at the auditory nerve level.

[25]The reader might query the validity of this 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 is, however, a simple and convenient procedure which 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 *et al.* did in their simulation.

[26]Damper *et al.* (submitted) deal with the reason for this, which is to do with the details of the synthesis strategy, at some length. To use their description, the bilabial stimuli are "pathological". It is interesting that the BSB model seems to be sensitive to this pathology in a different way – by producing too small a VOT value for the bilabial category boundary, see Figure 6(a). This is precisely the effect found by Damper *et al.* for a competitive-learning network trained using the Rumelhart and Zipser (1985) algorithm.

References

ABRAMSON, A., & LISKER, L. (1970). Discrimination along the voicing continuum: Cross-language tests. *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: Lawrence Erlbaum.

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

ANDERSON, J.A., SILVERSTEIN, J.W., & RITZ, S.A. (1977a). Vowel pre-processing with a neurally based model. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '77)*, Hartford, CN, pp. 265–269.

ANDERSON, J.A., SILVERSTEIN, J.W., RITZ, S.A., & JONES, R.S. (1977b). Distinctive features, categorical perception, and probability learning: Some applications for a neural model. *Psychological Review*, **84**, 413–451.

ANDERSON, J.A., & SILVERSTEIN, J.W. (1978). Reply to Grossberg. *Psychological Review*, **85**, 597–603.

ARBIB, M.A. (Ed.) (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.

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

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. (1997). A biocybernetic simulation of speech perception in humans and animals. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, Orlando, FL, **in press**.

DAMPER, R.I., HARNAD, S., & GORE, M.O. (submitted). A computational model of the perception of voicing in initial stops. *Journal of the Acoustical Society of America*.

DAMPER, R.I., PONT, M.J., & ELENIUS, K. (1990). Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus. *Speech Transmis-*

sion Laboratory Quarterly Progress and Status Report (Royal Institute of Technology (KTH), Stockholm), **STL-QPSR 4/1990**, 7–41.

DAMPER, R.I., PONT, M.J., & ELENIUS, K. (1996). Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus. In W.A. Ainsworth (Ed.), *Advances in speech, hearing and language processing, vol. 3 (part B)* (pp. 497–546). Greenwich, CT: JAI Press.

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, UK: 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. (1988). 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: Lawrence Erlbaum.

ELMAN, J.L., & ZIPSER, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, **83**, 1615–1626.

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., STEYVERS, M., & LARIMER, K. (1996). Categorical perception of novel dimensions. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 243–248). Hillsdale, NJ: Lawrence Erlbaum.

GREEN, D.M., & SWETS, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley,

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

GROSSBERG, S. (1968b). Some physiological and biological consequences of psychological postulates. *Proceedings of the National Academy of Sciences, USA*, **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, UK: Academic.

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

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. (Ed.) (1987). *Categorical perception: The groundwork of cognition*. Cambridge, UK: 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*, Stanford, CA, 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 cognit-*

ive modeling: Steps towards principled integration (pp. 191–206). London: Academic Press.

HARY, J.M., & MASSARO, D.M. (1982). Categorical results do not imply categorical perception. *Perception & Psychophysics*, **32**, 409–418.

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

HELSTON, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York, NY: Harper & Row.

HESSEN, A.J. VAN, & SCHOUTEN, M.E.H. (1992). Modeling phoneme perception. II: A model of stop consonant discrimination. *Journal of the Acoustical Society of America*, **92**, 1856–1868.

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

HOPFIELD, J.J. (1982). Neural networks and physical systems with emergent collective abilities. *Proceedings of the National Academy of Sciences, USA*, **79**, 2554–2558.

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.

KAPLAN, H.L., MACMILLAN, N.A., & CREELMAN, C.D. (1988). 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.

KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59–69.

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

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.

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. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967* (pp. 563–567). Prague: Academia.

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

LUCE, R.D. (1959). *Individual choice behavior*. New York, NY: John Wiley.

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, UK: 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, The Netherlands: Martinus Nijhoff.

MACMILLAN, N.A., & CREELMAN, C.D. (1991). *Detection theory: A user's guide*. New York, NY: 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. (1975). *Experimental psychology and information processing*. Chicago, IL: Rand McNally.

MASSARO, D.W. (1987a). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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.

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). Possible acoustic bases for the perception of voicing contrasts. In M.E.H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 188–198). Dordrecht, The Netherlands: Martinus Nijhoff.

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

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 time to comparisons within and across phonetic categories. *Perception & Psychophysics*, **15**, 285–290.

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

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.

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

REMEZ, R. (1987). Neural models of speech perception: A case history. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 199–225). Cambridge, UK: Cambridge University Press.

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

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, UK: Cambridge University Press.

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

ROSEN, S., & HOWELL, P. (1987). Auditory, articulatory and learning explanations of categorical perception of speech. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 113–160). Cambridge, UK: Cambridge University Press.

ROSNER, B.S. (1984). Perception of voice onset time: A signal-detection analysis. *Journal of the Acoustical Society of America*, **75**, 1231–1242.

RUMELHART, D.E., HINTON, G.E., & WILLIAMS, R.J. (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, Vol. 2: Psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

RUMELHART, D.E., & ZIPSER, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75–112.

SCHOUTEN, M.E.H. (1980). The case against a speech mode of perception. *Acta Psychologica*, **44**, 71–98.

SCHOUTEN, M.E.H., & HESSEN, A.J. VAN (1992). Modeling phoneme perception. I: Categorical perception. *Journal of the Acoustical Society of America*, **92**, 1841–1855.

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.

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.

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.

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.

WILSON, M. (1987). Brain mechanisms in categorical perception. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 387–417). Cambridge, UK: Cambridge University Press.

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.

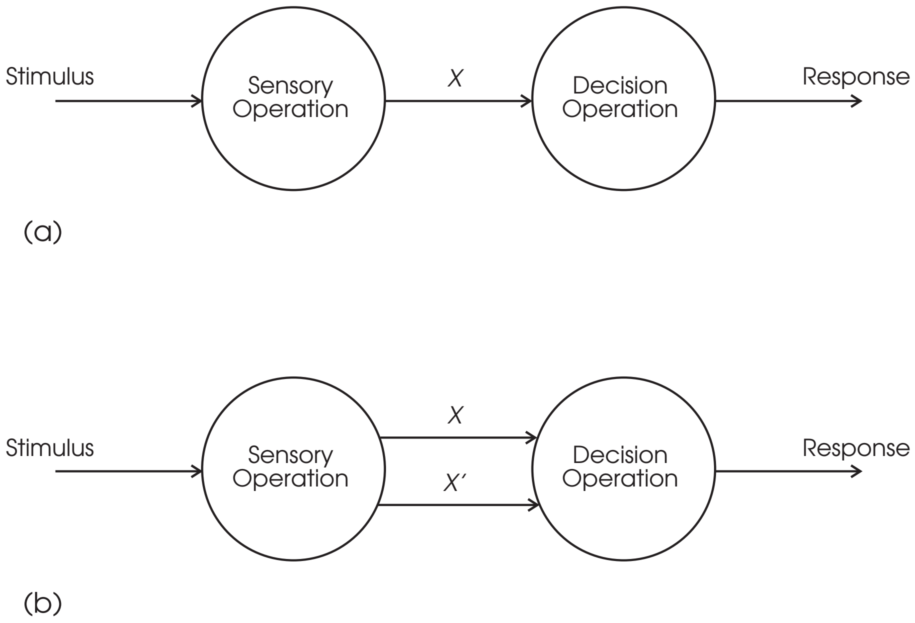
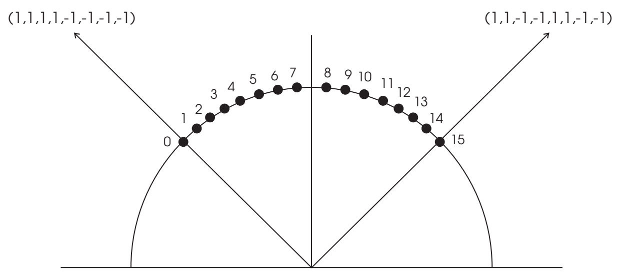
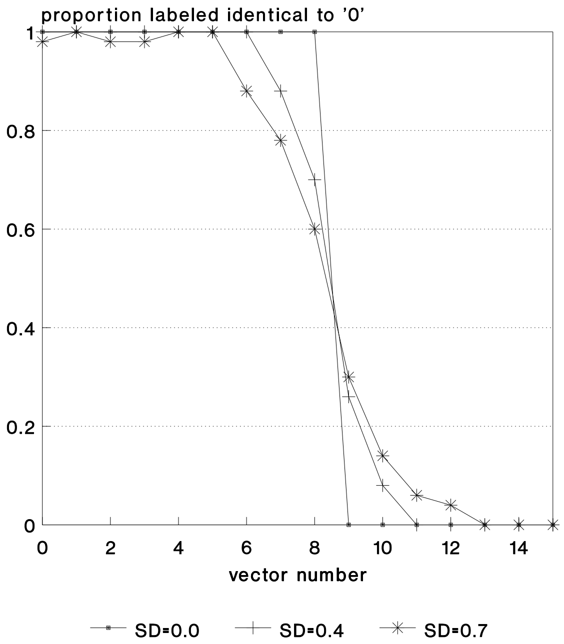


Figure 1: (a) The transformation of stimulus to response can be seen as a two-stage process of a sensory operation followed by a decision operation. This is consistent with signal detection theory's separation of sensitivity and response bias measures. *Redrawn from Massaro, 1987b*. (b) In Fujisaki and Kawashima's dual-process theory, there are two routes from sensory processing to decision: one continuous (X) and the other discrete (X').



(a)



(b)

Figure 2: (a) 8-dimensional system with two orthogonal length-2 eigenvectors used in the simulation of categorical perception based on the brain-state-in-a-box neural model. Inputs consisted of 16 equally-spaced length-1 vectors as shown, with added noise. *Redrawn from Anderson* et al., *1977b.* (b) Response of model for the simulated identification task with 1000 presentations at each noise condition: standard deviation 0.0, 0.4 or 0.7.

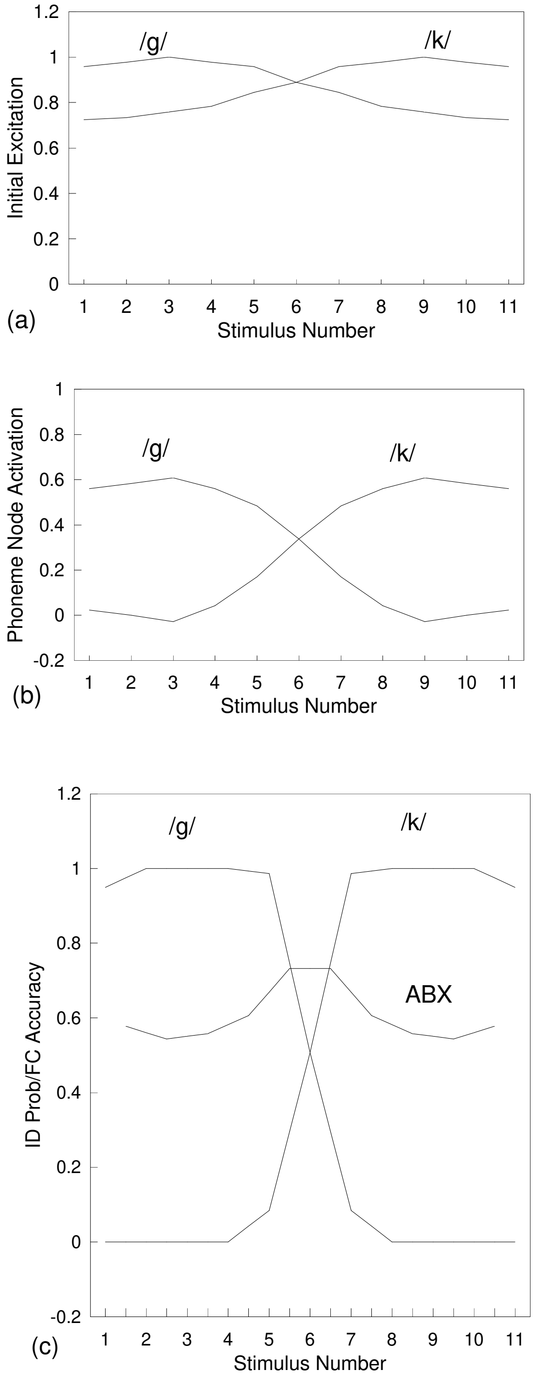
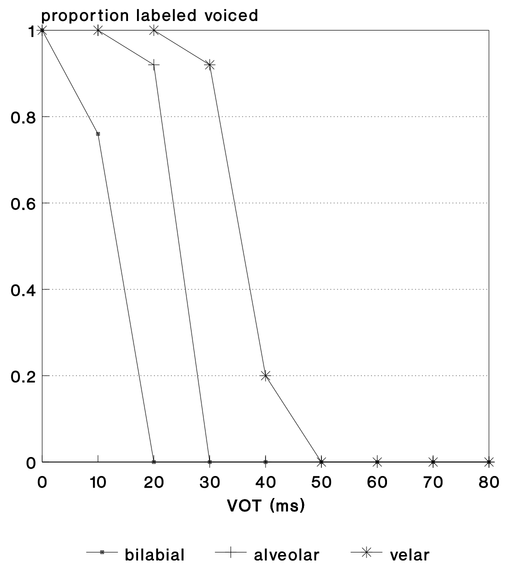
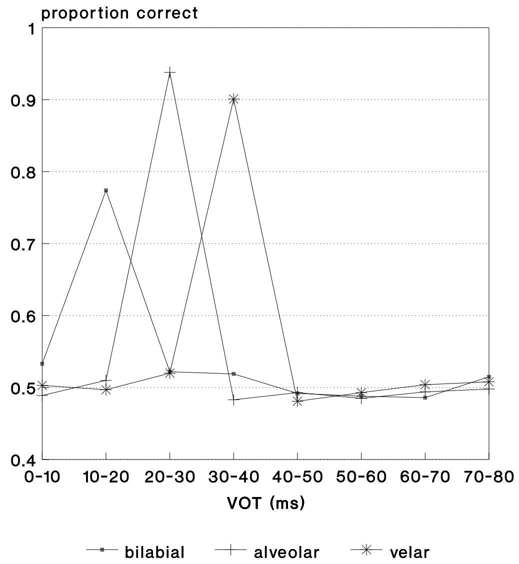


Figure 3: Categorical perception in the TRACE model: (a) Initial activation of the /g/ and /k/ units arising from bottom-up influence of the feature units, at time step $t = 1$. (b) Activations at time step $t = 60$. (c) Labeling functions after post-processing using Luce's choice model with $k = 5$, and *ABX* discrimination curve. *Redrawn from McClelland and Elman, 1986*.



(a)



(b)

Figure 4: Categorical perception of voice-onset time 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 1000 presentations.

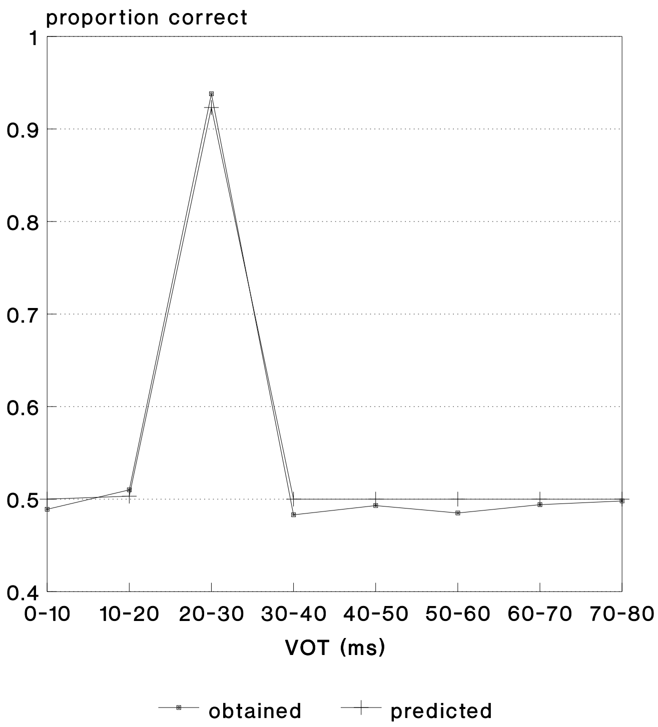
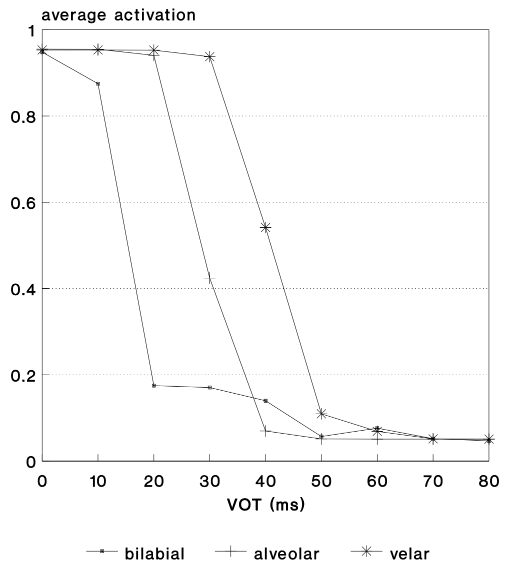
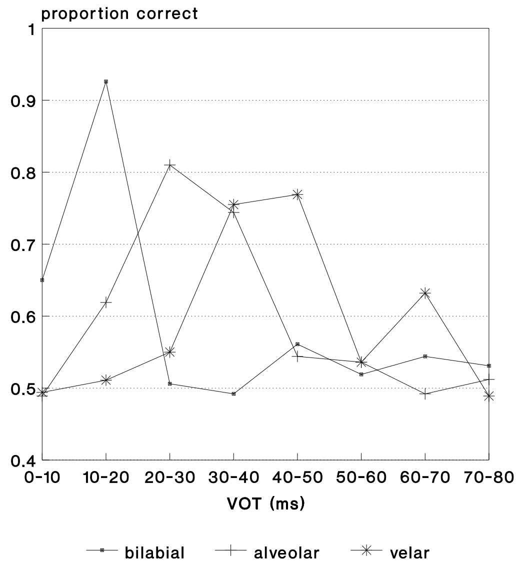


Figure 5: Categorical perception of voice-onset time in the brain-state-in-a-box model: Obtained *ABX* discrimination function for the alveolar series and that predicted on the basis of the corresponding labeling function using the Haskins formula. The close correspondence reflects the fact that the neural model converts the input to a discrete label, so that its discrimination function is a direct implementation of the Haskins formula.



(a)



(b)

Figure 6: Categorical perception of voice-onset time by multilayer perceptrons with 2 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 1000 presentations: 500 of $\langle ABA \rangle$ and 500 of $\langle ABB \rangle$. The guessing threshold g was 0.001.

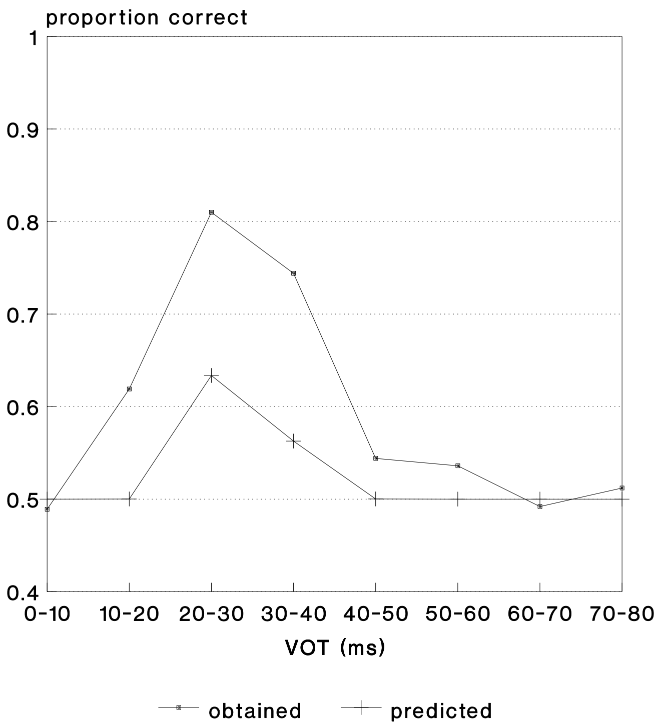


Figure 7: Categorical perception of voice-onset time by MLP with 2 hidden units: Obtained one-step *ABX* discrimination function for the alveolar series and that predicted on the basis of the corresponding labeling function using the Haskins formula. Here, the obtained discrimination is better than predicted (as seen in psychophysical results) reflecting the fact that the output of the neural model is continuous rather than discrete.

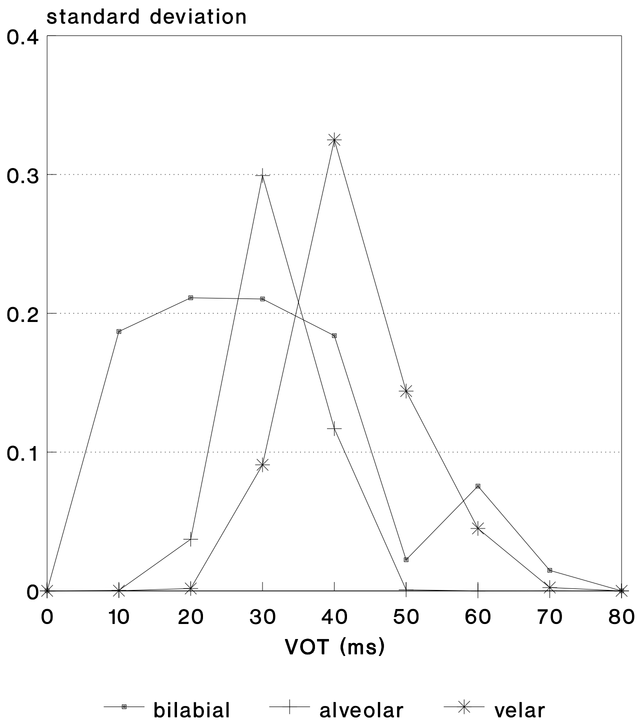


Figure 8: Standard deviation of the MLP's output activation. Each point is an average of 50 presentations.

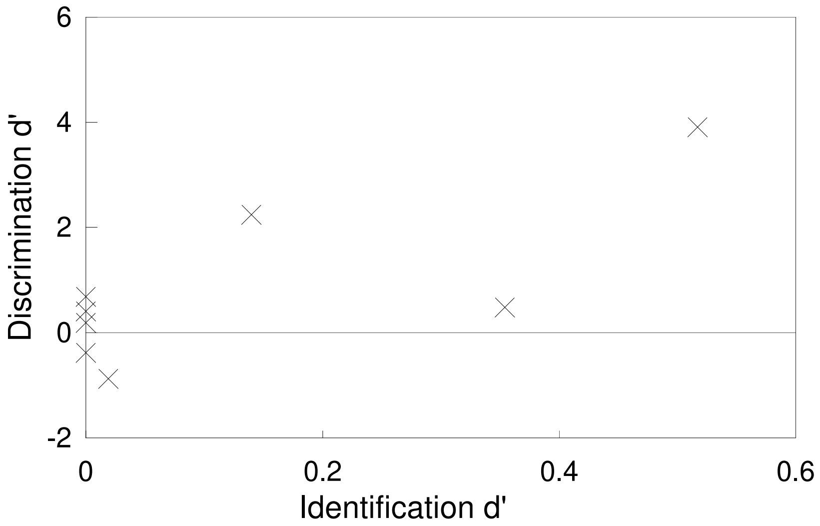


Figure 9: Discrimination d' versus identification d' for the back-propagation neural model and the alveolar VOT series. Categorical perception according to Macmillan *et al.* (1977) requires that these two sensitivity measures are equivalent, i.e. in the ideal case, all points should lie on the unity-slope line.

iterations	h_1h_2 coding		both different?
	0 ms	80 ms	
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

Table 1: Ranked number of iterations required to train the 192-2-1 bilabial multilayer perceptron to an error criterion of $\epsilon = 0.25$, starting from 10 different random, initial weight settings, and the hidden unit codings h_1h_2 discovered for the 0 ms and 80 ms VOT endpoints on which the net was trained.

Presentation	Response 1	Response 2
$\langle ABA \rangle$	hit	miss
$\langle ABB \rangle$	false alarm	correct rejection
$\langle BAA \rangle$	false alarm	correct rejection
$\langle BAB \rangle$	hit	miss

Table 2: Hits, misses, false alarms and correct rejections in the *ABX* discrimination paradigm. In the work reported here, the perfect 'memory' of the computer simulations means that the order of the *A* and *B* standards was irrelevant. Hence, only the stimulus-response matrix in the top half of the table was collected.

VOT (ms)	Identification proportion	Hit rate (H)	$d'_s = 2z(H)$	d'	Identification distance
0	0.955	–	–	–	–
10	0.955	0.4840	-0.0800	-0.380	0
20	0.941	0.8280	1.8960	2.246	0.014
30	0.424	0.9720	3.8200	3.910	0.517
40	0.070	0.5260	0.1306	0.480	0.354
50	0.051	0.4220	-0.3934	-0.873	0.019
60	0.051	0.4500	-0.2512	0.682	0
70	0.051	0.5180	0.0902	0.404	0
80	0.051	0.5040	0.0200	0.190	0

Table 3: SDT analysis of synthetic categorical perception by the back-propagation neural model: alveolar series.