

1994-12

Neural Dynamics of Variable Rate Speech Categorization

<https://hdl.handle.net/2144/2176>

"Downloaded from OpenBU. Boston University's institutional repository."

**NEURAL DYNAMICS OF VARIABLE-RATE
SPEECH CATEGORIZATION**

Stephen Grossberg, Ian Boardman, and Michael Cohen

December 1994

Revised: September 1995

Technical Report CAS/CNS-94-038

JEP: HPP, in press

Permission to copy without fee all or part of this material is granted provided that: 1. the copies are not made or distributed for direct commercial advantage, 2. the report title, author, document number, and release date appear, and notice is given that copying is by permission of the BOSTON UNIVERSITY CENTER FOR ADAPTIVE SYSTEMS AND DEPARTMENT OF COGNITIVE AND NEURAL SYSTEMS. To copy otherwise, or to republish, requires a fee and/or special permission.

Copyright © 1995

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA 02215

NEURAL DYNAMICS OF VARIABLE-RATE
SPEECH CATEGORIZATION

by

Stephen Grossberg‡, Ian Boardman†, and Michael Cohen‡

Department of Cognitive and Neural Systems
and

Center for Adaptive Systems
Boston University
111 Cummington Street
Boston, Massachusetts 02215 USA

Technical Report CAS/CNS-TR-94-038
Boston, MA: Boston University

December, 1994

Revised: September, 1995

Prepared for *Journal of Experimental Psychology:*
Human Perception and Performance, in press

Please address reprint inquiries to Stephen Grossberg.

‡ Supported in part by Air Force Office of Scientific Research (AFOSR F49620-92-J-0225).

† Supported in part by Advanced Research Projects Agency (ARPA AFOSR 90-0083), Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), and Pacific Sierra Research Corporation (PSR 91-6075-2)

The authors wish to thank Diana Meyers for her valuable assistance in the preparation of the manuscript.

ABSTRACT

What is the neural representation of a speech code as it evolves in real time? A neural model of this process, called the ARTPHONE model, is developed to quantitatively simulate data concerning segregation and integration of phonetic percepts, as exemplified by the problem of distinguishing "topic" from "top pick" in natural discourse. Psychoacoustic data concerning categorization of stop consonant pairs indicate that the closure time between syllable final (VC) and syllable initial (CV) transitions determines whether consonants are segregated, i.e., perceived as distinct, or integrated, i.e. fused into a single percept. Hearing two stops in a VC-CV pair that are phonetically the same, as in "top pick," requires about 150 msec more closure time than hearing two stops in a VC₁-C₂V pair that are phonetically different, as in "odd ball." When the distribution of closure intervals over trials is experimentally varied, subjects' decision boundaries between one-stop and two-stop percepts always occurred near the mean closure interval (Repp, 1980). The neural model traces these properties to dynamical interactions between a working memory for short-term storage of phonetic items and a list categorization network that groups, or chunks, sequences of the phonetic items in working memory. These interactions automatically adjust their processing rate to the speech rate via automatic gain control. The speech code in the model is a resonant wave that emerges after bottom-up signals from the working memory select list chunks which, in turn, read out top-down expectations that amplify consistent working memory items. The resonance between bottom-up and top-down information develops on a slower time scale than the processing of bottom-up information alone. It focuses attention upon speech groupings in working memory that are expected based upon past experience, while inhibiting speech features that are not expected, as in phonemic restoration. As in other examples drawn from Adaptive Resonance Theory, it is proposed that all conscious speech percepts are resonant events. In the case of VC₁-C₂V pairs, such a resonance may be rapidly reset by inputs, such as C₂, that are inconsistent with a top-down expectation, say of C₁; or, in the absence of a top-down mismatch, by a collapse of resonant activation due to a habituating process that can take a much longer time to occur, as illustrated by the categorical boundary between VCV and VC-CV. The categorical boundary for integration of VC-CV persists 150 msec longer than that of VC₁-C₂V because of the resonant dynamics that subserve perception of C. These categorization data may thus be understood as emergent properties of a resonant process that adjusts its dynamics to track the speech rate.

Key Words: speech perception, categorization, working memory, chunking, attention, neural network, adaptive resonance theory, ART, consciousness

1. Introduction: The Resonant Dynamics of Conscious Speech Percepts

What is the nature of the process that converts brain events into behavioral percepts? An answer to this question is needed to understand how the brain controls behavior, and how the brain is, in turn, shaped by environmental feedback that is experienced on the behavioral level. The nature of this connection also needs to be understood to develop neurally plausible connectionist models. Without it, a correct linking hypothesis cannot be developed between psychological data and the brain mechanisms from which it is generated.

The present article illustrates the hypothesis that conscious speech percepts are emergent properties that arise from resonant states of the brain. Such a resonance develops when bottom-up signals that are activated by environmental events interact with top-down expectations, or prototypes, that have been learned from prior experiences. The top-down expectations carry out a matching process that selects those combinations of bottom-up features which are consistent with the learned prototype, while inhibiting those that are not. In this way, an attentional focus starts to develop that concentrates processing on those feature clusters that are deemed important, based upon past experience. The attended feature clusters, in turn, reactivate the cycle of bottom-up and top-down signal exchange. This reciprocal exchange of signals eventually equilibrates in a resonant state that binds the attended features together into a coherent brain state. Such resonant states, rather than the activations due to bottom-up processing alone, are proposed to be the brain events that represent conscious behavior.

A classical example of such a matching process occurs during phonemic restoration (Samuel, 1981; Warren, 1984; Warren and Sherman, 1974). Suppose that a noise is followed immediately by the words "eel is on the ...". If that string of words is followed by the word "orange", then under proper temporal conditions, subjects hear "peel is on the orange". If the word "wagon" completes the sentence, "wheel is on the wagon" is heard. If the final word is "shoe", one hears "heel is on the shoe". Such experiences show that a bottom-up stimulus alone, such as "noise-eel", may not determine a conscious perception. Rather, the percept may be determined by the sound that one expects to hear in that auditory context, based on previous language experiences.

To explain such percepts, one needs to understand why "noise-eel" is not heard before the last word of the sentence is even presented. This may be explained by the fact that, if the resonance has not developed fully before the last word is presented, then this word can influence the expectations that determine the conscious percept. One also needs to explain how the expectation can convert "noise-eel" into a percept of "peel". This is attributed to the top-down matching process that selects expected feature clusters for attentive processing, while suppressing unexpected ones. In the "noise-eel" example, those spectral components of the noise are suppressed that are not part of the expected consonant sound.

This selection process directly influences phonetic percepts. It is not merely a process of symbolic inference. For example, if silence replaces noise, then only silence is heard. If a reduced set of spectral components is used in the noise, then a correspondingly degraded consonant sound is heard (Samuel, 1981).

Given that a resonant event may lag behind the environmental stimuli that cause it, one needs to develop a refined concept of how perceived psychological time is related to the times at which stimuli are presented. In particular, how can "future" events influence the perception of "past" events, yet time is perceived to always flow from past to future? The theory suggests how this is accomplished by a resonant wave that develops from past to future, even while it incorporates future constraints into its top-down decision process until each event in the resonance equilibrates.

In order to represent such a process, one needs to distinguish the external input rate from the internal rate at which the resonance process evolves. Since external events may, in principle, occur at arbitrary times, the internal rate process must have a finer time scale than any detectable external rate. It must also be faster than the resonance time scale that emerges

as a result of bottom-up and top-down interactions. That is why differential equations are used in the model. Differential equations are the universally accepted mathematical formalism in science that is used to describe events that are evolving in real time.

A related concern is: How can future events influence past events without smearing over all the events that intervene? In particular, how can silent intervals be perceived between the words “peel” and “orange” in “peel is on the orange” even after “orange” crosses all the intervening sounds to influence “peel”? Here again the nature of the top-down matching process is paramount. This matching process can *select* feature components that are consistent with its prototype, but it cannot “create something out of nothing”. Silence remains silence, no matter how active the top-down prototypes may be.

The opposite concern is also of importance. How can sharp word boundaries be perceived even if the sound spectrum that represents the words exhibits no silent intervals between them? The theory proposes that silence will be heard between words whenever there is a temporal break between the resonances that represent the individual words. In other words, silence is a discontinuity in the rate at which resonance evolves.

In order to make these concepts precise and workable, an analysis of psychological space, no less than of psychological time, is also required. In particular, it is not sufficient to posit processing levels that proceed, say, from letters to words, as in the popular Interactive Activation Model (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982); see Section 12. The language units that are familiar to us from daily experience, such as phonemes, letters, and words, do not form appropriate levels in a language processing hierarchy. Such a representation cannot learn stable representations of words in an unsupervised fashion, and is not consistent with various data about word recognition (Grossberg, 1984, 1986). Rather, processing levels that compute more abstract properties of auditory processing are needed, in particular, a working memory (Baddeley, 1986; Cohen and Grossberg, 1986; Grossberg, 1978a; Miller, 1956) is posited herein which represents sequences of “items” that have been unitized through prior learning experiences. Such items are familiar feature clusters that are presented within a brief time interval.

As items are processed through time, they generate an evolving spatial pattern of activation across the working memory. This spatial pattern represents both item information (which items are stored) and temporal order information (the order in which they are stored). A number of articles have modeled the design principles governing such item-and-order working memories and have used them to explain data about free recall (Grossberg, 1978a, 1978b), reaction time during sequential motor performance (Boardman and Bullock, 1991; Grossberg and Kuperstein, 1986/1989), errors in serial item and order recall due to rapid attention shifts (Grossberg and Stone, 1986a), errors and reaction times during lexical priming and episodic memory experiments (Grossberg and Stone, 1986b), and data concerning word superiority, phonemic restoration, and backward effects on speech perception (Grossberg, 1986). Such a wide range of data fall under the purview of these working memory models because they all satisfy two simple postulates (Bradski, Carpenter, and Grossberg, 1992, 1994; Grossberg, 1978a, 1978b). These postulates lead to working memories that can store sequences of events in a way that enables them to be grouped, or unitized, into categories, or “list chunks”, by a learning process that retains its stability even as new events are stored in the working memory through time.

In like manner, the working memory described in this article interacts with a categorization network that unitizes sequences of items by activating nodes that represent list chunks. These list chunks may represent the items themselves or larger groupings of items, such as phonemes, letters, syllables or words. The chunking network is designed to select those list categories that are most predictive of the temporal context that the items, taken together, collectively generate across the working memory as its activity pattern evolves through time (Cohen and Grossberg, 1986, 1987; Grossberg, 1978a, 1986; Grossberg and Stone, 1986a, 1986b). Such chunking networks have been used to explain a variety of data about catego-

rization, including the Magic Number Seven (Grossberg, 1978a) of Miller (1956).

As the most predictive chunks are selected through competitive interactions, they read out the learned top-down prototypes that are matched against the items in working memory. This is how the contextually correct item groupings are selected, including the groupings that replace “noise-eel” in the phonemic restoration example that was discussed above. Thus, by closing the bottom-up top-down feedback loop, the model clarifies how the process of unitization is linked to the process of phonetic perception.

Learning plays a key role in rationalizing why brain resonances exist that bind features into attentional states. These resonant dynamics are modeled by Adaptive Resonance Theory, or ART, mechanisms that were introduced in Grossberg (1976a, 1976b; see also Grossberg (1980a) for an early review). ART proposes that top-down matching focuses attention so that the brain can rapidly learn new information without just as rapidly being forced to forget previously learned information that is still useful. In other words, ART shows how the brain learns to solve the *stability-plasticity dilemma* (see Carpenter and Grossberg (1991, 1992, 1993) and Grossberg (1995) for recent reviews). ART learning hereby escapes the type of catastrophic forgetting that bedevils all feedforward learning models, including the popular back propagation model of Werbos (1974) and Parker (1982) that was popularized within the cognitive science community by Rumelhart, Hinton, and Williams (1986).

A number of previous articles have been devoted to showing how ART mechanisms can be used to explain speech and language data (Bradski, Carpenter, and Grossberg, 1992, 1994; Cohen and Grossberg, 1986, 1987; Grossberg, 1978a, 1978b, 1986; Grossberg and Stone, 1986a, 1986b). The present article analyses data concerning category boundary shifts that are measured when VC₁-C₂V pairs are experienced. Repp (1980) presented /ib/-/ga/ and /ib/-/ba/ syllables to subjects under conditions that are described more fully in Section 3. In brief, the silence interval was varied between the initial vowel-consonant syllable and the terminal consonant-vowel syllable. If the silence was short enough, then /ib/-/ga/ sounded like /iga/ and /ib/-/ba/ sounded like /iba/. Repp (1980) showed that the transition from perceiving /iba/ to /ib/-/ba/ requires around 150 milliseconds more silence than the transition from /iga/ to /ib/-/ga/. One hundred fifty milliseconds in a very long time compared with the time needed to activate neurons, which is at least an order of magnitude smaller. Why is this shift so large? We trace it below to how the /ib/-/ga/ and /ib/-/ba/ resonances evolve through time. In particular, a mismatch between /g/ and the internal representation of a recently presented /b/ can reset the resonance that /b/ would otherwise have generated, leading to a switch from an /ib/-/ga/ percept to an /iga/ percept at relatively short silence intervals. On the other hand, a second /b/ can prolong the resonance due to a recently presented /b/, thus allowing the percept /iba/ to supplant /ib/-/ba/ until much longer silence intervals.

A related issue about the processing of psychological time also needs to be understood to explain these data. If a resonance can lag behind the stimulus event that cause it, then why do not resonances take so long to occur that brain events cannot keep up with the rate at which stimuli are presented? This problem could become acute during processes like speech perception which need to respond to both slow and fast rates of input presentation. We show below how a process of automatic gain control can track the speech rate through time and use this running estimate to speed up or slow down the processing rate in the working memory and chunking network. As a result, the delays at which resonances emerge and the times at which they terminate can keep up with the speech rate. In fact, finer properties of the Repp (1980) data show categorical boundary shifts that are sensitive to the mean silence interval; that is, the “speech rate” in his paradigm. We show that the *variability* of these category boundaries derives from the system’s efforts to generate a speech code that is *invariant* under changes in the speech rate.

In a similar fashion, during phonemic restoration, the maximum duration of the noise intervals that permits an uninterrupted speech percept is nearly equal to the average duration

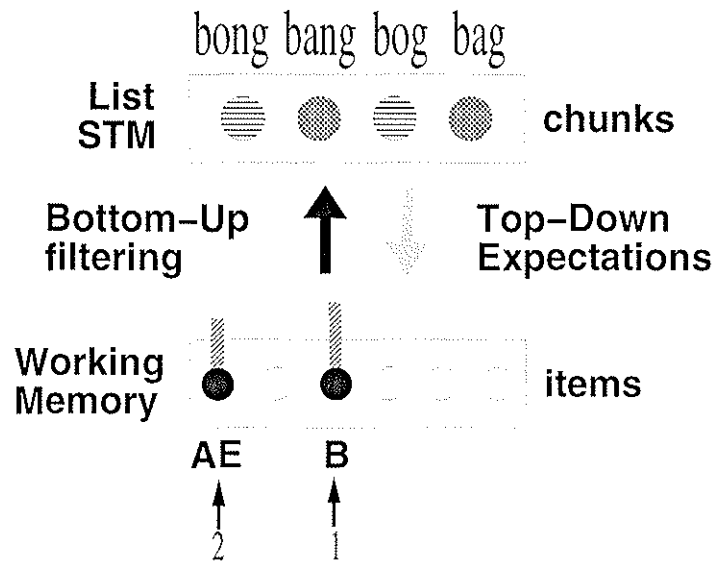


Figure 1. A working memory that stores phonemic items interacts with list categories through bottom-up and top-down adaptive filters. Item lists in working memory prime the list categories, which in turn send top-down expectation signals back to the working memory to reorganize its contents through a matching process.

of the most frequent units in the speech stream (Bashford and Warren, 1987). Thus, noise intervals may be roughly as long as the average syllable duration when disconnected syllables are presented. A similar effect was reported by Repp, Liberman, Eccardt, and Pesetsky (1978). They presented the words “gray chip” using an interval of fricative noise in place of the second word’s initial consonant, and noted that “gray chip” can be heard as “great ship” merely by increasing the duration of the noise. The /t/ percept appears to be labile, and may either group with the /f/ (“sh”) to form the affricate /f/ (“ch”), or move across the intervening silence, grouping with the syllable that preceded the noise to form a word. In this study too, the temporal boundary shifted with the average speaking rate of the carrier sentence. These and other studies discussed below indicate that the interactive grouping process is modulated by contextually determined timing information, resulting in percepts that are invariant with variable global speech rate.

The discussion which follows explains how sensitivity to temporal variations can be incorporated in the reciprocal interactions between a working memory and a chunking field to produce rate-invariant speech percepts. The dynamics of matching input against expectation provide an account of temporal integration and segregation of phonetic percepts. This leads to the development of a neural network model of the interactive feedback process. Computer simulations of the model closely approximate human subject performance in discriminating stop-consonant pairs. The model is also compared with alternative models for explaining speech and language data, in particular the fuzzy logical model of perception (Massaro, 1989), the interactive activation model (McClelland and Rumelhart, 1981), and the TRACE model (McClelland and Elman, 1986).

2. Adaptive Resonance in Speech and Language Processing

As noted above, the dynamical interaction between items in working memory and list chunks is illustrative of a cyclical process that has been described by Adaptive Resonance Theory, or ART (Carpenter and Grossberg, 1991; Grossberg, 1978a, 1980a, 1986). The present application of ART to model phonetic percepts is called the ARTPHONE model.

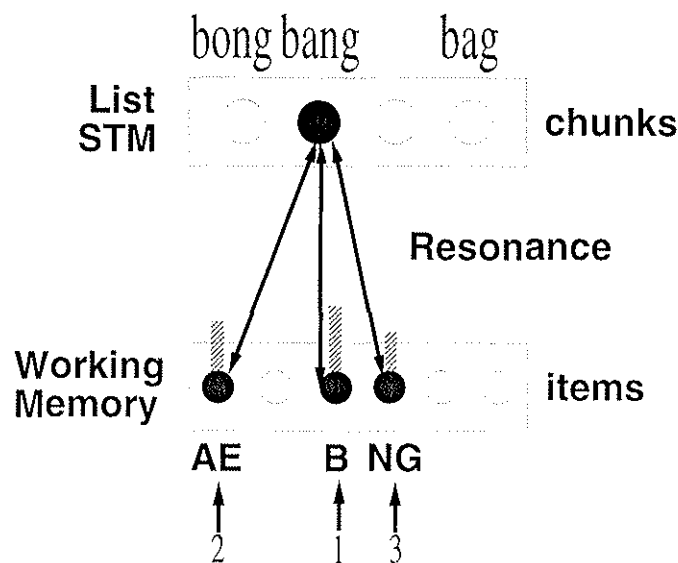


Figure 2. A particular list category wins the competition at the chunk level and generates top-down excitatory feedback that represents the category's expectation. Matching between bottom-up list items and the top-down expectation selects those item features that are consistent with the expectation and suppresses the rest. ("AE" is "a" as in "hat", "NG" as in "bong").

An earlier version of the model was briefly reported in Boardman, Cohen, and Grossberg (1993). Within this model, the interaction at the phonemic processing stage begins with the speech signal being preprocessed (via prior stages of adaptive resonance) into unitized item representations. These items are sequentially stored in a working memory (Baddeley, 1986; Bradski, *et al.*, 1992; Grossberg, 1978a, 1978b; Grossberg and Stone, 1986). The rate of processing in the working memory automatically adjusts itself based on temporal information in the speech signal so that the speech representation remains approximately invariant even with variable long-term speech rates.

As items enter the memory buffer, working memory item nodes send bottom-up priming signals to list chunk nodes via adaptive filters, activating several potential item groupings, or list categories (see Figure 1). For example, as a spoken word beginning with "ba..." (/bæ/) enters the working memory, it sequentially activates populations responsive to the /b/ and then the /æ/. These items prime chunks encoding lists that start with /b/ and /bæ/. List chunks exhibit properties of self-similarity, so that chunks for longer lists require greater input to exceed threshold (Cohen and Grossberg, 1986; Grossberg, 1978a). Furthermore, larger chunks inhibit or *mask* smaller ones, so that larger lists containing a prescribed sublist are favored over smaller ones. See Grossberg (1984) for a discussion of relevant data.

Thus the /bæ/ chunk can become fully activated only after the /æ/ item is activated. As the /bæ/ chunk becomes active, it suppresses the /b/ chunk. Once active, list chunks begin to send top-down feedback to associated items. These top-down signals represent a learned expectation of the pattern that is stored in working memory (Figure 2). Those chunks whose top-down signals are best matched to the sequence of incoming data reinforce the working memory items and receive greater bottom-up signals from them. Mismatched chunks are either not activated in the first place, or are progressively inhibited by recurrent inhibition from the better matched chunks. As the best matched chunks receive excitatory signals from and emit excitatory signals to the working memory, they continue to reinforce one another. As a result, a resonant wave travels across the network that embodies the speech code and percept. For example, completion of the word "bang", as in Figure 2, extends the pattern in

working memory, matching the expectation from the list category for “bang” and reinforcing it. The emerging resonance enables this category (and possibly related sublist categories) to win the competition at the chunk level, and to complete the recognition event.

The resonant process can be interrupted or terminated by two different mechanisms: *mismatch reset* and *habituated collapse*. Mismatch reset is due to mismatch between the top-down expectations and incoming bottom-up data. When an input pattern arises in working memory that fails to match an active category’s top-down expectation, the category loses its bottom-up support while simultaneously being suppressed by competition from other categories that make a better match with the input pattern. Mismatch reset has already been used to model many other speech and language data, including reaction time and error data about lexical priming and decision processing (Grossberg and Stone, 1986b; Schvaneveldt and McDonald, 1981).

Habituated collapse can occur after a resonance develops and a category maintains resonant activation levels for some time. The synaptic transmitter in the excitatory pathways between the list category and its associated working memory items gradually become inactivated or habituated (Grossberg, 1986, Section 28). When habituation progresses past a certain point, the signals in the pathways can no longer support the resonance. Activation decays below the signal threshold, and a category “collapse” occurs. Reset due to habituated synaptic transmitters has also been used to model many other brain phenomena, including visual persistence (Francis, Grossberg, and Mingolla, 1994), visual afterimages (Francis and Grossberg, 1995a; Grossberg, 1976a), form-motion interactions (Francis and Grossberg, 1995b), binocular vision (Grossberg and Grunewald, 1995), circadian rhythms (Carpenter and Grossberg, 1983, 1984, 1985) and the control of arm movements (Gaudio and Grossberg, 1991).

In summary, a resonance can either self-terminate after fully unfolding and habituating its transmitters, or it can be actively reset by an input mismatch, possibly even before reaching resonant levels of activation. Both cases will be illustrated in the simulated data. More generally, all the key elements of the model – its working memory, chunking network, matching and resonance rules, and habituated transmitters – have been used to explain many other behavioral and brain data. In this sense, the ARTPHONE model provides a parsimonious explanation of the data targeted herein by using basic model mechanisms that seem to be utilized in many brain systems.

The model is elaborated below after a review of perceptual phenomena that it will be used to simulate. The focus is on the grouping over time of categorical responses, often described as the temporal integration and segregation of phonetic percepts (Repp, 1988). Integration maps multiple speech segments, e.g., phones, onto a single mental unit that unifies them into a single percept. Segregation maps multiple segments onto distinct mental units or percepts. Processes in the model perform these perceptual functions through dynamical feedback interactions between item and list categories that are proposed to represent speech percepts.

3. Segregation and Integration of Stop Consonant Percepts

An example of a phonetic grouping phenomenon is stop consonant gemination. Here, *gemination* refers to the percept of a double consonant that can arise from a single closure production. In English, production of a stop consonant embedded between two vowels could be perceived either as a single stop, within a word, or as two stops across a word boundary; e.g., “topic” vs. “top pick.” Gemination is generally cued by longer closure duration, but can also be signaled by the burst at the onset of the second consonant. There is a temporal boundary, called the *single-geminate* boundary, at which one or two stops are equally probable percepts. Pickett and Decker (1960) showed that this boundary, typically around 200 msec, was sensitive to the global speech rate context. In Italian, where double stops can appear within some words, the boundary is also sensitive to temporal cues, in

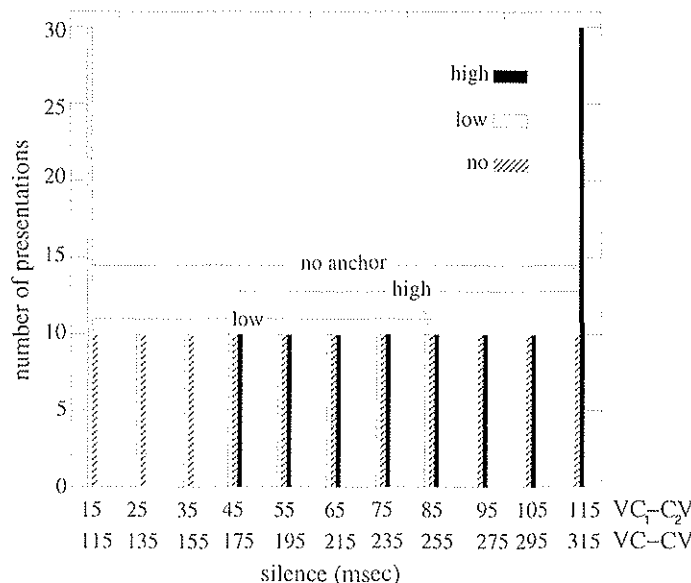


Figure 3. Distributions of silent intervals used for $VC_1 - C_2V$ (upper time scale) and $VC - CV$ (lower time scale) stimuli used in Repp (1980).

particular, the duration of the preceding vowel (Rossetti, 1994).

English phonotactic rules permit two phonetically different stop consonants (e.g., /b/ and /d/) to appear consecutively only when the first ends a syllable and the second begins the next syllable. The closure interval can be artificially shortened, however, to the point that the syllable-final consonant may not be perceived by the listener. That closure duration for which one or two stops are equally probable percepts defines a *single-cluster* boundary (Dorman, Raphael, and Liberman, 1979; Repp, 1978). This boundary was reported to be approximately 70 msec.

Repp (1980) continued his investigation of the role of the closure duration in integration and segregation of stop consonant percepts. The purpose of the Repp (1980) experiment was to determine how single-cluster and single-geminate boundaries respond to changes in the long-term statistics (mean and variance) of silent intervals across trials. It was not expected that the single-cluster boundary would be sensitive to this manipulation, because it was thought that the interference due to later occurring formant transitions reflected an essentially acoustic or *pre-categorical* processing (Repp, 1978). The study failed to confirm this expectation and, instead, provided quantitative information about the adaptation of categorical percepts to long-term speech rate.

The Repp (1980) experiments used stimuli consisting of pairs of vowel-consonant (VC) or consonant-vowel (CV) syllables. A VC syllable was always followed by a CV syllable. These syllable pairs were separated by a silent closure interval of variable duration. There were two sets of experiments. In one set, the two consonants were perceived as phonetically distinct, while in the other set, they were perceived as the same. The consonants were in all cases the voiced stops /b/ or /g/.

A phonetically different pair of stops is referred to as a stop *cluster*. A phonetically similar pair of stops is called *geminate* if it produces a double stop percept. The duration of closure for each trial was chosen randomly from a set, according to one of three probability distributions for each of the two classes of stop pairs (see Figure 3). The “no anchor” case covered the full range of silence intervals, a set of 11 specific values equally spaced across the range, with uniform probability of being presented. The “low anchor” case used a subset of the 8 shortest intervals with a higher probability of the shortest interval. The “high

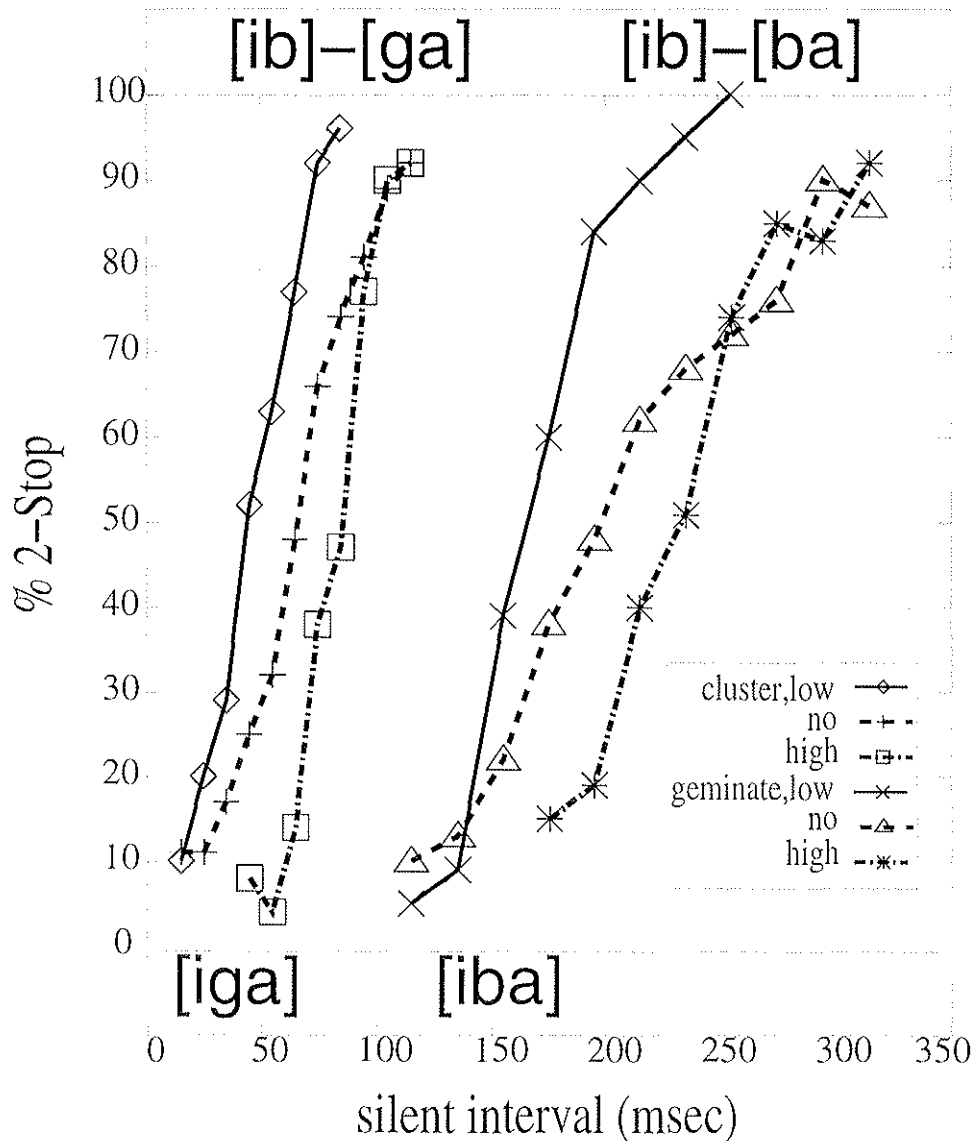


Figure 4. Psychometric functions in response to the 2-syllable stimuli for all 6 conditions, averaged over 8 subjects. (Data reprinted with permission from Repp, 1980, Figures 1 and 2.)

anchor” case used a subset of the 8 longest intervals with a higher probability of the longest interval. Depending on the silent interval presented, the VC_1-C_2V stimulus is perceived by the subject as either VC_2V or VC_1-C_2V ; and the $VC-CV$ stimulus is perceived as either VCV or $VC-CV$. With three ranges of silence intervals for both $VC-CV$ and VC_1-C_2V stimuli types, there were a total of six experimental conditions.

Eight subjects participated, including Repp himself. All had previous exposure to synthetic speech sounds. The results, averaged over all subjects, are shown in Figure 4. Repp (1980) reported that “all subjects tested showed these shifts, including the author who, despite foreknowledge and to his considerable surprise, was affected just as much as the other subjects.” The curves describe category boundaries between one and two stops for each experimental condition. For *all* conditions, subjects were more likely to perceive two stops as the silent interval between the two syllables increased. The horizontal shift of the curves in

relation to the range of silence interval used in each condition indicate that the subjects' decisions were strongly influenced by the distribution of silences across trials. In fact, averaged over subjects, the decision boundary appears to be established right around the mean silent interval used in that condition (see the value of the abscissa at the 50% probability point of each curve in Figure 4.) Also clearly evident is the broad time gap between the single-cluster and single-geminate boundaries. Hearing two of the same stop typically requires about 150 msec more closure time than does hearing two different stops.

The shift of category boundaries can be viewed from two perspectives. On the one hand, a two-stop stimulus with a given interval can be perceived as a single stop if it arises in a series for which the mean silent interval is long, but as two stops if the mean silent interval is sufficiently short. On the other hand, the same percept can be obtained from a range of silent intervals if the ratio of the silent interval presented to the mean silent interval is fixed. For example, the subject has a 50% probability of hearing two stops whenever the silent interval in the stimulus equals the mean silent interval for the series. Thus the percept can be said to remain *invariant* with changes in the long term average silent interval.

Concerning the statistical significance of these results, Repp (1980) wrote: "One obvious further point to consider is the possibility that the large range-frequency effects were simply a consequence of the large region of uncertainty, reflected in the shallow slopes of the identification functions. This question was investigated by computing product-moment correlations across subjects between slopes of individual identification functions and extent of boundary shift. While a negative correlation of -0.55 was found in the single-geminate condition (which supports the hypothesis that smaller slopes go with larger shifts), a positive correlation of +0.59 was found in the single-cluster condition (which contradicts the hypothesis); both correlations were nonsignificant. It should also be noted that individual identification functions were often considerably steeper than the average functions shown in Figure 4, and there were several instances of large boundary shifts despite steep slopes. Thus, no convincing evidence for a direct relation between uncertainty and sensitivity to range-frequency was found within the present experiment, suggesting that a shallow-sloped identification function is neither necessary nor sufficient for large context effects to occur."

Repp (1980) replicated these results with a second experiment that used natural speech stimuli, rather than synthetic speech sounds. A two-alternative forced choice discrimination task was used to study the single cluster condition. Here, the first stop consonant, that preceded /g/, was varied to be either /b/ or /d/. Again a range-frequency effect on category boundaries was found, as in Figure 4.

4. Description of Phonetic Grouping in the Model

The model provides an integrated account of the gap between cluster and geminate conditions and boundary shifts reported in Repp (1980) as follows. First, assume a consonant is consciously perceived only when resonance between item and list categories raises the consonant's category node above some output threshold (see Figure 5A). When two inputs which each activate different item nodes are presented in rapid succession, the response of the category node, or nodes, associated with the first item will be reset due to category mismatch with the second item. This may occur before the category node activation has exceeded the output threshold. Hence the response to the first input would be undetected, since only suprathreshold resonances lead to conscious percepts (Figure 5B). Even if the activation does exceed the output threshold, it can still be rapidly reset if a mismatch occurs with respect to a subsequent input (Carpenter and Grossberg, 1991; Grossberg, 1986; Grossberg and Stone, 1986b).

When two inputs which both activate the same item node are presented sufficiently close in time, the integrated responses may "fuse" into a single suprathreshold event (Figure 6A). This can happen because the second activation of the item node can occur while the resonance due to its first activation is still intact. Without an intervening sub-threshold interval, the

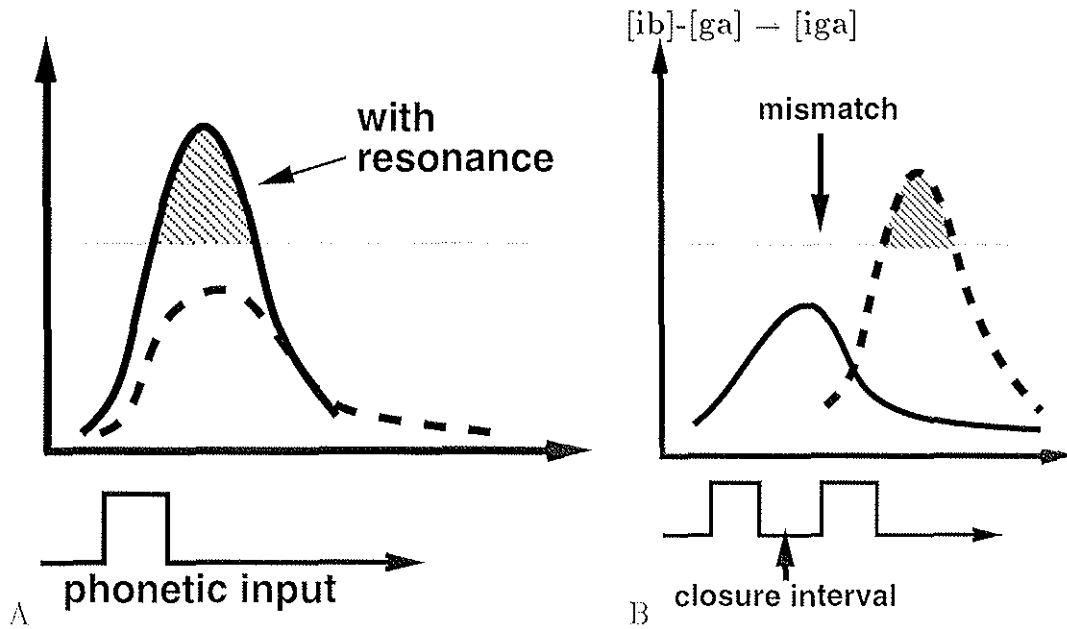


Figure 5. (A) Response to a single stop with (solid line) and without (dashed line) resonance. The ordinate represents category node activity and the abscissa represents time. Suprathreshold activation (above horizontal line) is shaded. (B) Reset due to phonologic mismatch. Here the activity corresponding to /b/ is reset by mismatch with /g/ before it can resonate. Only the /g/ sound reaches resonance, leading to a percept of /iga/.

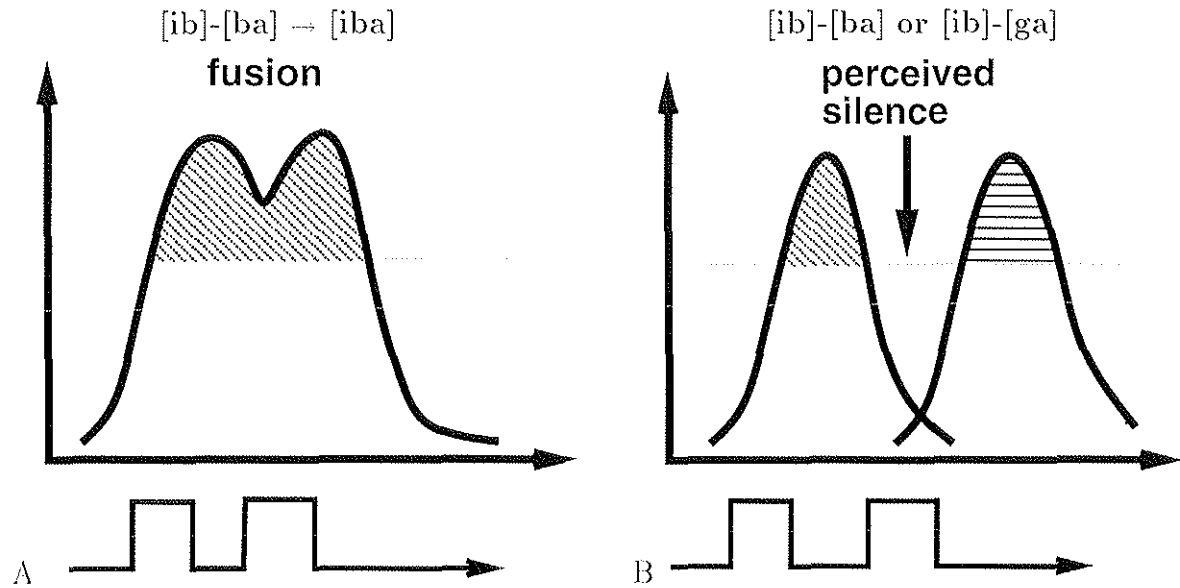


Figure 6. (A) Fusion in response to similar iterated /b/ phones leads to a prolongation of the /iba/ resonance through time. (B) A sufficiently long silent interval allows a 2-stop percept to be heard. Habituated collapse of the /ib/ resonance before the /ba/ or /ga/ resonance develop leads to a percept of both the VC and CV sounds.

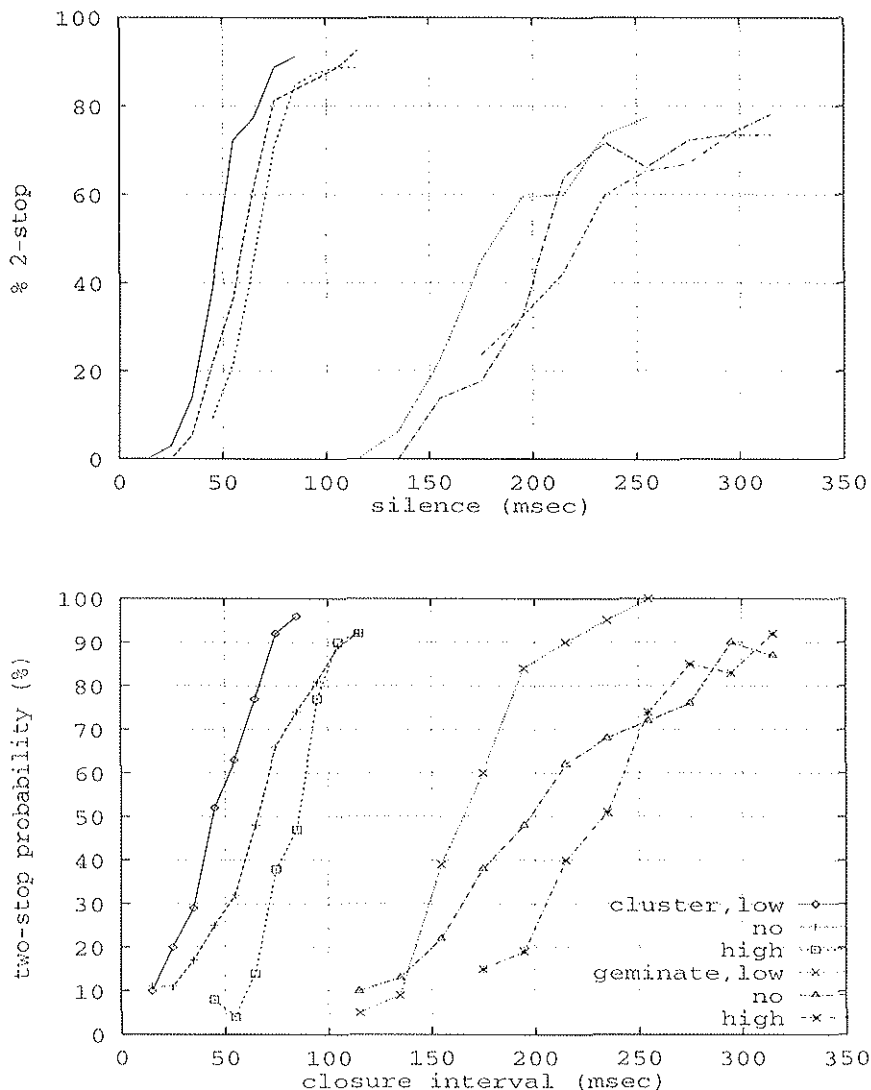


Figure 7. Top graphs plot computer simulations, bottom graphs the original data of Repp (1980); see Figure 4. Parameters: $\alpha = 0.5$, $\beta = 1$, $\gamma = 0.097$, $\delta = 0.28$, $\kappa = 100$, $\chi = 1$, $\phi = 5 \times 10^{-4}$, $\eta = 6$, $\tau = 350$, $\nu = 0.005$, $\zeta = 0.11$, $\lambda = 0.1$, $\mu = 2.2$, $\sigma = 0.06$, $\theta = 0.22$.

system can detect only one prolonged resonance in response to the two inputs. Sufficiently long silence following a given input allows a resonant response to terminate due to habituated collapse. When its activity falls below threshold, the associated percept ends and begins an interval of perceived silence (Figure 6B). A second percept of the /b/ sound can thus develop when /b/ is presented as part of /ba/. This example illustrates how resonant timing may reorganize the time scale of external events to define discontinuous gaps in the rate at which a resonant wave evolves, and thus a perception of silence.

Figures 7 and 8 summarize ARTPHONE simulations of these category boundaries and compare them with the Repp (1980) data. Unlike other models of speech categorization, which typically plot category boundaries as a function of the two alternative percepts (Massaro, 1989; Massaro and Cohen, 1993; Massaro and Oden, 1995; McClelland, 1991; McClel-

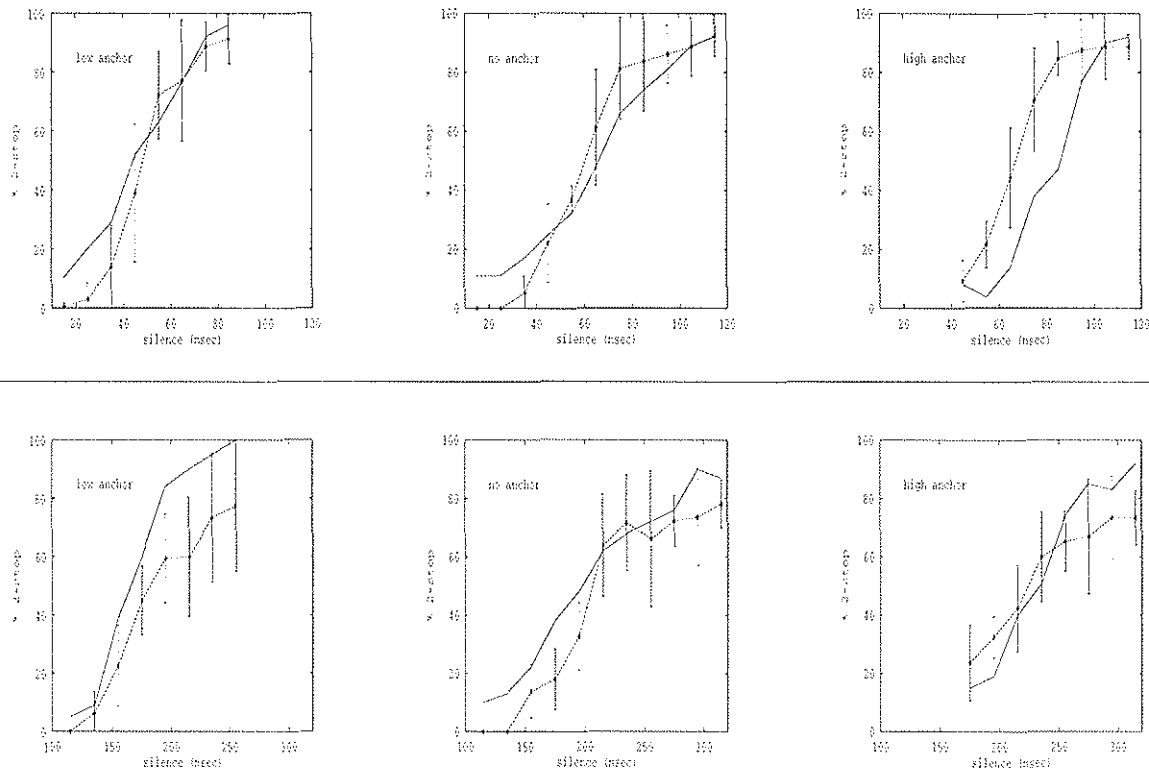


Figure 8. Repp (1980) data, thin solid lines; computer simulation, lines with points. Top panel: cluster case. Bottom panel: geminate case. Each simulation curve is an average over 8 runs of 100 trials each. Error bars represent 1 standard deviation from the mean.

land and Elman, 1986), the ARTPHONE model computes category boundaries as they are created from their emerging speech representations in real time. The subsequent sections describe the model in detail and show how these representations and their category boundaries are formed.

5. ARTPHONE MODEL

The model has been implemented as a neural network representing a working memory for phonetic items and a list category stage. The dynamics of resonance, category mismatch and collapse are governed by differential equations to represent the continuous unfolding of system dynamics in real time. These properties are demonstrated in the computer simulations presented below. The simulations show how the model can transform item sequences in working memory into categorical responses which shift in time with changes in mean input rate.

It cannot be overemphasized that the category boundaries simulated in this way do not represent a curve fit in the sense that this concept is usually understood in fitting psychological data. The model does not fit the data to pre-established functions that represent data curves with some free parameters. Rather, the model generates category boundaries as an *emergent property* of the system-wide interactions that give rise to its resonances. The model's most important function is to dynamically generate internal representations of the data through its bottom-up and top-down interactions. The properties of these representations are then fit to the data.

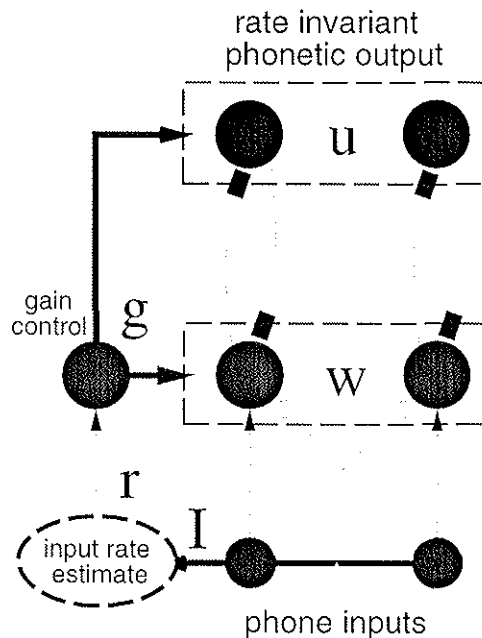


Figure 9. Working memory item activities (w) excite list chunk activities (u) through previously learned bottom-up pathways. List chunk activities send top-down excitatory feedback down to their item source nodes. Bottom-up and top-down pathways are modulated by habituated transmitter gates (filled squares). Item nodes receive input in a on-center off-surround anatomy. Total input (I) is averaged to control an item rate signal (r) that adjusts the working memory gain (g). This gain tracks the speech rate and adjusts the integration rates of working memory and chunking network accordingly. Excitatory paths are marked with arrowheads, inhibitory paths with small open circles.

There are no pre-established curves against which to fit the data. Thus, the number of *parameters* that are needed to define model interactions is not the key factor in determining model complexity. Rather, it is the number of *processes* that are needed to explain the data, and whether the qualitative properties of each of these processes are robust under parameter changes that do not unbalance the system. In the present instance, only two processes, a working memory and a chunking network are needed. Both of these processing levels have been implicated in numerous other explanations of speech, language, and motor control data. In this sense, the model is parsimonious and even elementary. On the other hand, a grand unified simulation of all these data using one set of parameters remains a goal for future research at the present stage of model development.

The ARTPHONE model is shown schematically in Figure 9. The item nodes in the working memory layer encode partially compressed representations of the acoustic features of the speech sounds (Cohen, Grossberg and Stork, 1988; Grossberg, 1978a; Grossberg and Stone, 1986). The encoding is the result of learning by the adaptive weights, or long term memory (LTM) traces, that exist in the adaptive pathways from the acoustic feature representations to the item nodes. As incoming speech segments associated with words sequentially activate these tuned item nodes, spatial patterns of activation evolve across the working memory. Repeated exposure to specific spatial patterns permits learning by the LTM traces in the adaptive pathways between the item nodes and the list nodes. Just as any given activity pattern arises sequentially from smaller patterns, there are list nodes that categorize familiar sublists of any given list, even one-item lists known as *singletons*. For the purposes of the following discussion, we can assume that the adaptive tuning of the pathways that activate the

item and category nodes occurred during a critical developmental period and is stable; see Carpenter and Grossberg (1991, 1993) and Grossberg (1980a, 1986) for a discussion of model mechanisms with these properties. Thus, no adaptive weights or processes are included in the present implementation.

After a speech segment activates an item node, the item node then excites associated list nodes. These list nodes, in turn, activate excitatory top-down feedback to the item nodes, corresponding to learned expectations. This reciprocal exchange of bottom-up and top-down signals enables a resonant state to develop. Both the bottom-up and top-down excitatory signals pass through transmitter gates that are inactivated, or habituated, by the signals in their respective pathways (Carpenter and Grossberg, 1990; Grossberg, 1972, 1980a). When a transmitter is sufficiently habituated by the signals passing through its pathway, then the resonance supported by that pathway can begin to collapse, after which node activations decay passively.

Resonant activation of item nodes results from a combination of bottom-up input and top-down feedback. Any one item node may receive top-down signals from many list chunks as the input sequence progresses. The bottom-up input has an on-center off-surround organization, reflecting auditory anatomy at several levels of organization (Irvine, 1986; Pickles, 1988). Due to this anatomy, inhibition arising from subsequent inputs serves to suppress prior activations that mismatch the evolving top-down expectation. A second mechanism of mismatch reset via an orienting, or novelty-sensitive, subsystem (Carpenter and Grossberg, 1991; Grossberg, 1980a) is implemented in Section 11.

6. Rate Invariance and Gain Control

The time needed for resonance to bring activation to a perceptually significant level and for the resonant response to collapse is dependent on the neural activation rates in the network model. Activation rates play a critical role in adapting to speech rate by speeding up processing in response to a high rate of speech units and slowing it down in response to a slow rate. The parameter that controls an equations processing rate is called its *gain*. In the ARTPHONE model, the gain is automatically adjusted to a running estimate of the input rate (*cf.*, Grossberg, 1986, Section 45). Whether we hear “topic” or “top pick,” given a fixed target stimulus, is determined by the speech rate cue from the surrounding context (Pickett and Decker, 1960). In fact, the Repp (1980) data show that subjects’ percepts are approximately invariant with respect to mean silent interval between stop consonant clusters, which is inversely related to the input rate. These data suggest that an estimate of mean input rate may serve as a basis for adjusting the activation rate of the system so that phonetic percepts become independent of changes in the mean input rate.

7. Network Equations

With the basic concepts of the model described, we can next specify mathematically how these features are implemented. In the following network equations, Greek letters are fixed parameters. Each equation describes the time rate of change of a system variable x , denoted by $\frac{dx}{dt}$, in terms of its inputs and internal processes.

7.1 Item Working Memory Level

Let w_j be the activity of the j^{th} item representation p_j and I_j be its input. Then w_j obeys the equation

$$\frac{dw_j}{dt} = g(r)[(\beta - w_j)(I_j + u_j z_{ju}) - w_j(\alpha + \kappa \sum_{k \neq j} I_k)]. \quad (1)$$

In (1), term $g(r)$ represents the automatic gain control process. It multiplies the entire right hand side of (1) and thereby speeds up or slow down the rate $\frac{dw_j}{dt}$ with which w_j changes as it increases or decreases, respectively, through time. The gain term $g(r)$ is defined by

$$g(r) = \chi + \phi r^\eta. \quad (2)$$

In (2), χ is a constant, or tonic, baseline activation rate. Term ϕr^η is a variable rate, where ϕ is a constant and r adapts slowly to the input rate, as described in equation (7).

Term $(\beta - w_j)(I_j + u_j z_{ju})$ in (1) defines the excitatory effects of the bottom-up input I_j and the net top-down feedback signal $u_j z_{ju}$ on w_j . The top-down feedback signal u_j arises from the activity of the list chunk that is associated with item p_j via item node j . Signal u_j is multiplied, or gated, by the transmitter z_{ju} , which habituates in response to intense resonant activation of u_j , as shown in equation (5) below.

Both the bottom-up and top-down signals are multiplied by the shunting, or membrane equation, term $\beta - w_j$. This term assures that w_j cannot exceed β . It also causes the processing rate to be input-dependent, since all terms that multiply w_j on the right-hand side of (1) determine w_j 's net processing rate.

The inhibitory terms in (1) are combined in the expression $-w_j(\alpha + \kappa \sum_{k \neq j} I_k)$. Parameter α is a passive decay parameter. Multiplying the inhibitory input $-\kappa \sum_{k \neq j} I_k$ by the shunting term w_j assures that w_j never becomes negative. It also makes the rate $\frac{dw_j}{dt}$ of w_j processing dependent on all the terms I_k , $k \neq j$, which inhibit w_j via lateral inhibitory interactions.

Taken together, the excitatory and inhibitory interactions in (1) define a shunting on-center off-surround network. The on-center off-surround interpretation can be summarized as follows: The excitatory "on-center" input I_j can turn on only those $\beta - w_j$ cell sites that are unexcited, as in term $(\beta - w_j)I_j$. The inhibitory "off-surround" input $\sum_{k \neq j} I_k$ from other phones p_k , $k \neq j$, can turn off only those sites w_j that are already excited, as in term $-w_j \sum_{k \neq j} I_k$. Such a network has been proved capable of accurately detecting and storing the ratios $I_j(\sum_k I_k)^{-1}$ of the item inputs in working memory, assuming that they are simultaneously presented, even if the total input $\sum_k I_k$ becomes very large (Grossberg, 1973, 1980a). Such a shunting network extracts a normalized response even from wildly fluctuating inputs, while processing relative input importance without saturation.

7.2 List Chunking Level

As noted above, u_j is the activity of the list chunk associated with item node j . Activity u_j also obeys a gain-controlled shunting equation; namely,

$$\frac{du_j}{dt} = g(r)[(\beta - u_j)w_j^+ z_{jw} - \delta u_j]. \quad (3)$$

In (3), the gain control term $g(r)$ scales the rate $\frac{du_j}{dt}$ with which u_j changes, just as it did for w_j in (1). Also as in (1), only unexcited sites $(\beta - u_j)$ can be excited, here only by a bottom-up signal $w_j^+ z_{jw}$ from item j . This signal is derived from the activity w_j of item j via the thresholded signal $w_j^+ = \max(w_j - \alpha, 0)$. In particular, item j cannot begin to excite chunk j until its activity w_j exceeds threshold α . Once this signal is emitted from item node j , it is multiplied, or gated, by a transmitter z_{jw} which habituates in response to item

activity, as shown in (4) below. Thus both the bottom-up and the top-down transmitters between items and chunks habituate in response to activity in the pathways that they gate.

7.3 Transmitter Dynamics

The bottom-up transmitter z_{ju} and the top-down transmitter z_{jw} between the j^{th} item and category both obey the same habituating law, albeit in response to different signals. This law was introduced in Grossberg (1968, 1969). In the present instance, it becomes

$$\frac{dz_{jw}}{dt} = \zeta(1 - z_{jw}) - h(w_j^+)z_{jw} \quad (4)$$

and

$$\frac{dz_{ju}}{dt} = \zeta(1 - z_{ju}) - h(u_j)z_{ju}. \quad (5)$$

Consider equation (4) for definiteness. By term $\zeta(1 - z_{jw})$, transmitter accumulates to a target level 1 at a fixed rate ζ . When the system is at rest (namely $h = 0$ for a sufficiently long time), the transmitter variable equilibrates at its maximum value 1. By term $-h(u_j)z_{jw}$, transmitter is inactivated, or habituates, by a mass action interaction between the amount z_{jw} of available transmitter and a function $h(w_j^+)$ of the bottom-up signal w_j^+ emitted by item j . Thus z_{jw} is inactivated, and a resonant collapse initiated, at a rate proportional to how active item j becomes. Equation (5) says that the same is true for the top-down transmitter z_{ju} , except here the habituation rate increases as a function of $h(u_j)$, which grows with activity u_j of chunk j .

7.4 Transmitter Inactivation Rate

The transmitter inactivation rate is a nonlinear function of its signal, namely

$$h(x) = \lambda x + \mu x^2. \quad (6)$$

As first simulated in Gaudiano and Grossberg (1991), the higher-order term in (6) causes the gated signals wz_{jw} and uz_{ju} in (1) and (3) to exhibit non-monotonic responses as a function of the signals w and u , respectively, that cause their activity-dependent inactivation. To see this, solve (4) at equilibrium ($\frac{dz_{jw}}{dt} = 0$) to find that

$$z_{jw} = \frac{\zeta}{\zeta + h(w_j^+)}. \quad (7)$$

By (3), (6), and (7), the bottom-up signal from item j to list chunk j is then

$$w_j^+ z_{jw} = \frac{\zeta w_j^+}{\zeta + \lambda w_j^+ + \mu (w_j^+)^2}. \quad (8)$$

Thus, as activity w_j increases due to resonance, the signal (8) first increases and then decreases. The inactivation rule (6) hereby ensures that input signals eventually decrease at sufficiently high activation levels. The effect of varying parameters λ and μ in (6) is demonstrated in the simulations below.

8. Estimation of Input Density via Automatic Gain Control

It remains to define the automatic gain control $g(r)$ that modulates the integration rate of w_j and u_j in (1) and (3), respectively. This function tracks input density; namely, a time average of input energy. Time averaging of inputs is a simple leaky integration operation common to neurons, and is thus an intuitive mechanism for estimating the speech input rate. While Repp (1980) directly varied the mean silent interval between input speech segments, he consequently also varied the mean input density, because the number and amplitude of the input segments were fixed for the entire experiment. In the model, it is assumed that gain control is based on a running average of input density. We infer from the fact that decision boundaries are positioned near the mean (intra-stimulus) silent interval that subjects completely discount the much longer (2.5 second) intervals between stimulus presentations. Thus the estimator should ignore long intervals without input. These intervals can be discounted by limiting the input averaging to a fixed time frame, or window, following input. In the simulations that follow, the rate signal, r that goes into $g(r)$ is given by

$$\frac{dr}{dt} = w(t)(-\nu r + I), \quad (9)$$

where $I = \sum_k I_k$ is the total input,

$$w(t) = \begin{cases} 1 & I \geq \epsilon \\ 1 & t - t_c < \tau \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

and t_c is the last time that $I \geq \epsilon$ in a given input bout. By (9) and (10), r time-averages input I at rate ν and decays when input is not present, such that integration is gated off by w at τ msec after I falls below the threshold value ϵ . Whenever the input was on, it exceeded threshold ϵ in (10). The dependence of the average value of r on mean silent interval, s , can be interpolated by a decaying exponential function of the form.

$$r(s) = a_1 e^{-s/s_0} + a_2 \quad (11)$$

where the parameters a_1 , a_2 and s_0 are determined empirically assuming mean input energy and parameters ν and τ are fixed. For example, in the simulations below with $\nu = .005$ and $\tau = 350$ msec., $a_1 = 1.9$, $a_2 = 1.25$ and $s_0 = 256$ msec.

9. Computer Simulations of Variable-Rate Categorization

This section presents computer simulations of the model defined above. The simulations demonstrate the following properties of the model.

- An exchange of bottom-up and top-down feedback generates a resonance between phonetic items and list categories which is required for auditory perception.
- Category collapse arises following an interrupted resonance due to habituation of transmitter gates in the pathways between the item and list levels. The resonant phase that precedes collapse explains the long interval needed to segregate geminate stops.
- Category mismatch explains the short interval needed to distinguish between two stops in a consonant cluster.
- Changes in the network integration rate that track an on-line estimate of input rate explain shifts of psychometric functions in the cluster and geminate conditions.

All simulations were performed by numerical integration using a fourth-order Runge-Kutta method, with step size fixed at 0.1. At each time step, the simulator sets input

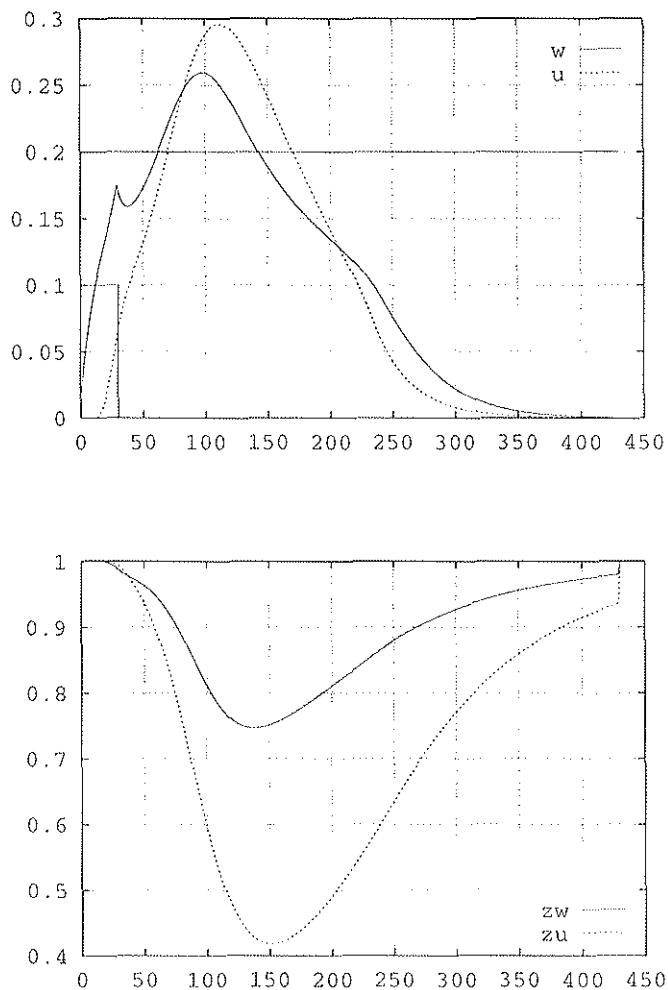


Figure 10. Activation time course for network variables, w , u (top), and z_w and z_u (bottom) through time. Rectangular pulse at lower left of top panel indicates the input interval and amplitude. Parameters: $\alpha = 0.5$, $\beta = 1$, $\gamma = 0.1$, $\delta = 0.28$, $\zeta = 0.1$, $\lambda = 0.1$, $\mu = 2$. Gain g was set to 1.1, a mid-range value. Time was calibrated in milliseconds.

variables according to an event schedule set by the experimenter, then sequentially computes the next change in each system variable while holding all values constant

9.1 Dynamics of Resonant Feedback

The first simulation demonstrates the response of the network to a single phone input. The graphs in Figure 8 show the time course of the network activities w and u in the upper panel, and of transmitters z_w and z_u in the lower panel. The indexing subscript j is redundant in this case. The network variables start at their rest levels: zero for activations w and u , unity for the z transmitter levels. When input has driven the item activity w above its output threshold γ , then a signal w^+z_w , as in (3), begins to activate u and list activity u begins to grow. Typically, item activity w begins to decay at the offset of input (see cusp in the w graph shortly after input terminates) because top-down list feedback from u is not

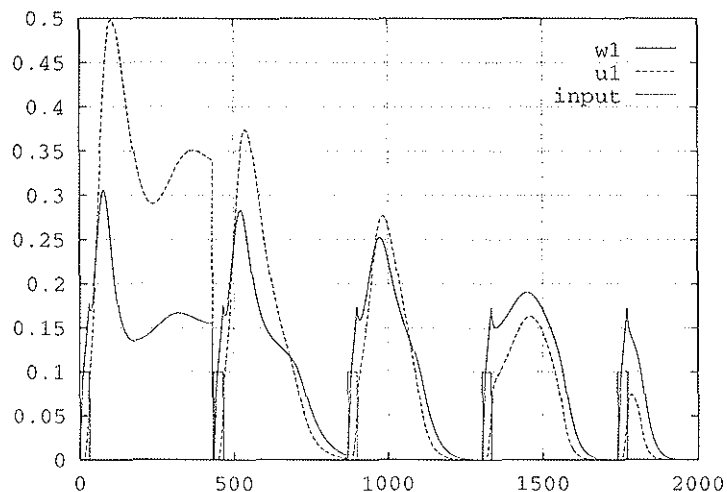


Figure 11. A series of five trials each showing w responses (solid curves) and u responses (dashed curves) to a single input pulse (rectangular graphs), as the passive decay rate δ in equation (3) increases. From left to right, $\delta = 0.1, 0.2, 0.3, 0.4, 0.5$. Other parameters are chosen as in Figure 10. Gain g was set to 1.2.

yet great enough to overcome the passive decay.

Bottom-up w excitation to the chunking node continues, however, at a level great enough to overcome the passive decay of u , even though w may be momentarily decaying, and so u continues to increase. Consequently, w changes from decreasing to increasing, an event called a resonant boost (ca. 40 msec in Figure 10A). Without this boost from top-down feedback, w and u would decay passively to rest. With top-down feedback, u increases until it exceeds a resonance detection threshold θ and thereafter reaches its resonance asymptote. During this phase, transmitter habituates, or is inactivated (Figure 10B), until eventually the gated excitatory input to the item activity w in (1) is no longer sufficient to support further growth, as when $uz_u < \alpha w/(\beta - w)$. List activity u then follows item activity w downward, resulting in category collapse.

An important rate inequality between systems (1) and (3) ensures the desired qualitative dynamics. The rate difference is established by constraining passive decay rates in (1) and (3) to satisfy $\delta < \alpha$. Consequently, list activity u follows its excitatory input w at a slower rate. The ability of top-down feedback from u to w to support w following termination of input is also consistent with the constraint that w decays rapidly relative to u . Without feedback from u to w , u is deprived of the extra excitatory boost in w input that it needs to exceed the resonance threshold.

The rate difference between (1) and (3) is chosen long enough to reflect the observed value of the single-cluster boundary, which in the model is determined by the time needed for u to exceed threshold θ (see Section 9.3 below). The lag is also constrained to avoid oscillations between u and w , which, for a given value of α , sets a lower bound for δ . Figure 11 shows the effect of varying δ . Oscillations, evident in the left-most trial of Figure 11 ($\delta = 0.1$), can arise when small passive decay δ permits list activity u to remain sufficiently large to retard the decline of item activity w even after substantial transmitter habituation occurs.

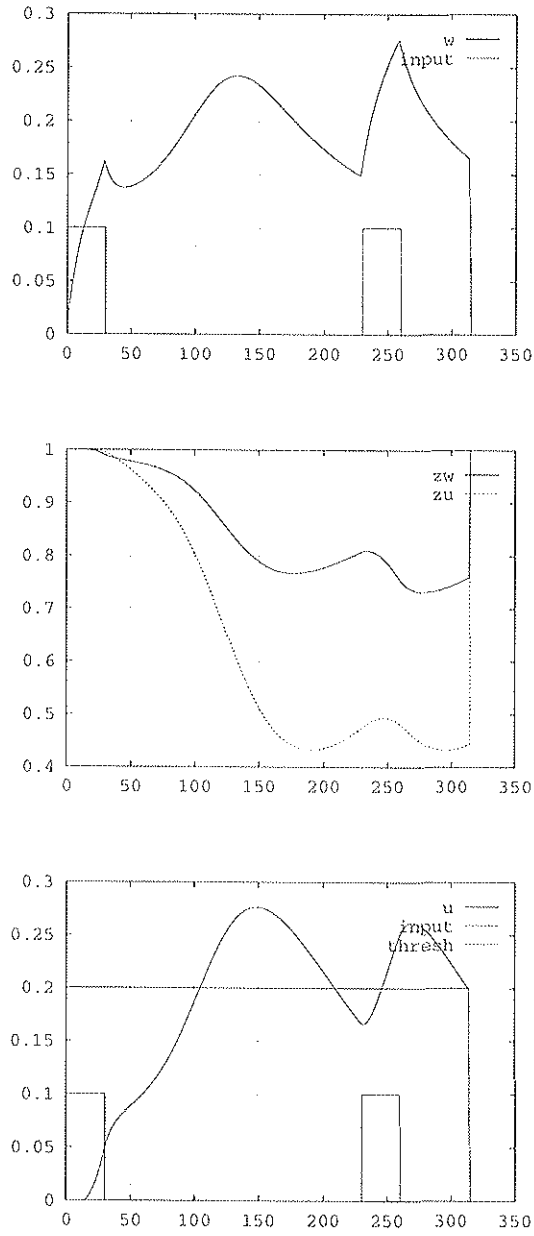


Figure 12. Activations w (top panel), z_w and z_u (middle panel), and u (lower panel) in response to two phonemically related inputs (rectangular pulses). The resonance threshold value u for perception is $\theta = 0.2$. Both are detected (see two suprathreshold u peaks).

9.2 Phonetic Segregation after Category Collapse

Any point in time that list activity u falls below the resonance threshold θ is assumed to begin an interval of perceived silence. Such a negative threshold crossing or *offset time* determines the earliest time that the next phonemically related input could be detected as separate phone. Thus it corresponds to the single-geminate boundary. Segregation of a pair of phonemically related inputs into separate category responses is demonstrated in the next simulation (see Figure 12).

The time interval when $u < \theta$ between the two suprathreshold phases of u models a silent interval that permits the detection of two distinct speech sounds.

9.3 Reset due to Mismatch

At any point in time before the list activity u reaches θ , the item activity w can be suppressed by a competing input, such as a /g/ instead of a /b/, that results in the collapse of u . In that case, the corresponding list category may not reach resonance and is thus not detected or “perceived” by the network. The time taken to achieve a positive going threshold crossing, or *onset time*, thus determines the minimum interval needed for phonetic perception, and corresponds to the single-cluster boundary. Reset of a list category’s response by a later occurring, different item is demonstrated in the next simulation (see Figure 13). This example demonstrates the case where only the second phone is perceived.

9.4 Transmitter Habituation

The time course of the resonant response is affected by the transmitter habituation parameters λ and μ in (6). Figure 14 shows two cases, one for which the nonlinear habituation parameter μ is zero and the other for which the linear habituation parameter λ is zero. The second-order μ term is more robust in producing a dynamic reset, as illustrated by equation (8). In fact, the u response continues to decay below threshold even as μ is reduced by more than half. By comparison, the response exhibits premature recovery when λ is reduced by a smaller proportion.

In general, increasing transmitter habituation lowers the range of the u response above threshold. Changes in transmitter habituation can thus change the onset and offset times that predict the single-cluster and single-geminate boundaries, and also the difference or gap between them. This gap has an additional dependence on the rate-controlled system gain $g(r)$ in (1) and (3), and transmitter habituation affects this dependence as well. To depict these relationships, we plot onset and offset times as a function of gain, and vary transmitter habituation, as shown in Figure 15. Increasing μ in (6) causes the onset and offset boundaries to move closer together, as expected from Figure 14B, but in addition, causes the gap between them to narrow at low values of gain. In other words, the u range above threshold is more sensitive to changes in mean input rate as μ increases.

9.5 Performance of Input Rate Estimator

The input rate signal r reflects a running estimate of the input density as given by (9) and (10). Each simulation of a Repp (1980) experiment begins with $r = 0$, and a fixed parameter ν in (9) is chosen so that r is within $\frac{1}{\epsilon}$ of asymptote by the end of ten warm-up trials (see Figure 16A). A more detailed view of the estimator is shown in Figure 16B, where r is shown for a set of discontinuous, randomly selected trials, one for each silent interval used in the simulation. Activation r increases during each input interval (always 30 msec), and decays passively following inputs. To save unnecessary computation, the simulator runs each trial only until the output response (u_j) following the second input has decayed below threshold. At that time, which is always less than the window time τ , variable r is set to the value it would have decayed to if the simulation had run till the window closed. Other network values are also set to their rest conditions. Going from left to right in Figure 16B,

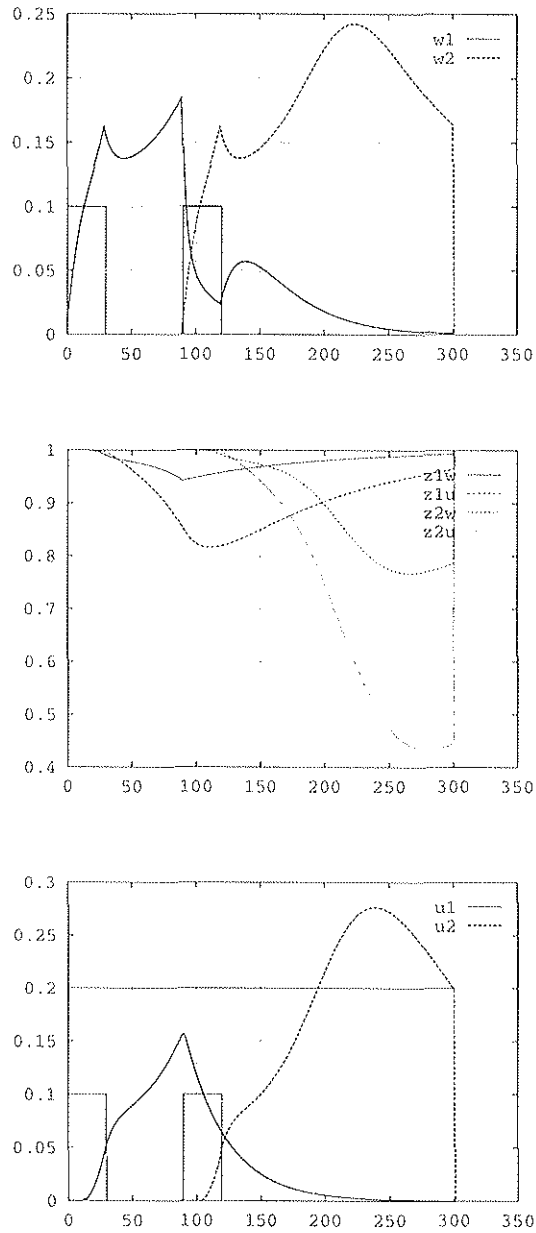


Figure 13. Activations w_j (top panel), z_{jw} and z_{ju} (middle panel), and u_j (lower panel) for a sequence of two phonemically unrelated inputs (rectangular pulses). Only the latter is detected (see the single suprathreshold u peak).

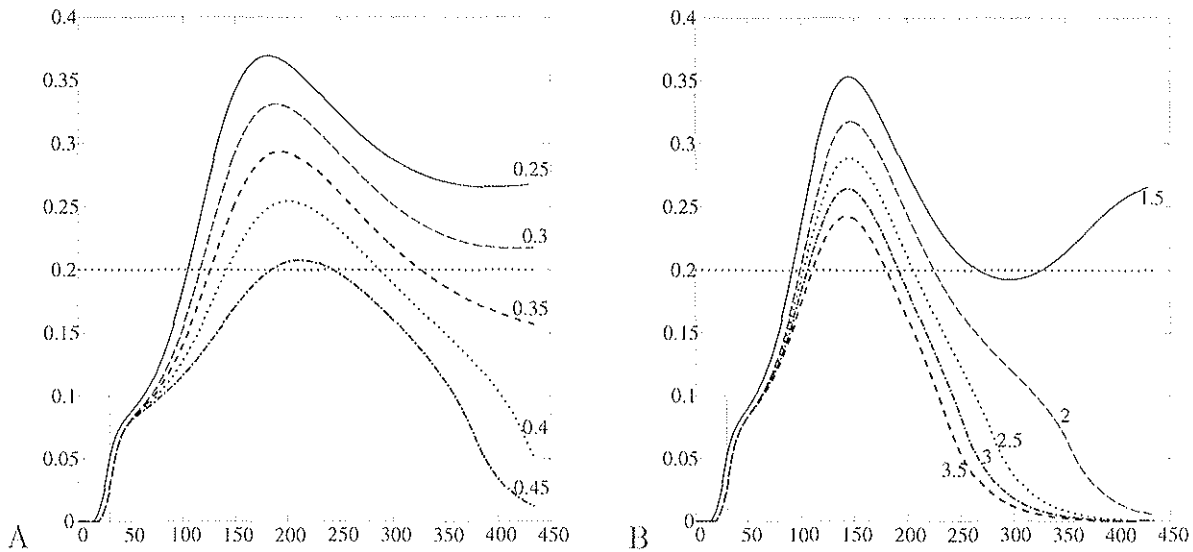


Figure 14. Category responses u in (3) through time to a single input for two cases of transmitter habituation: either (A) there is no second-order term ($\mu = 0$) in (6), with λ indicated along each curve; or (B) there is no linear term ($\lambda = 0$) in (6), with μ indicated along each curve. Other parameters are the same as in Figure 10. The resonance threshold $\theta = .2$ is indicated by the horizontal dotted line. The rectangular pulse in the lower left-hand corner is the input. Time along the abscissa is in milliseconds.

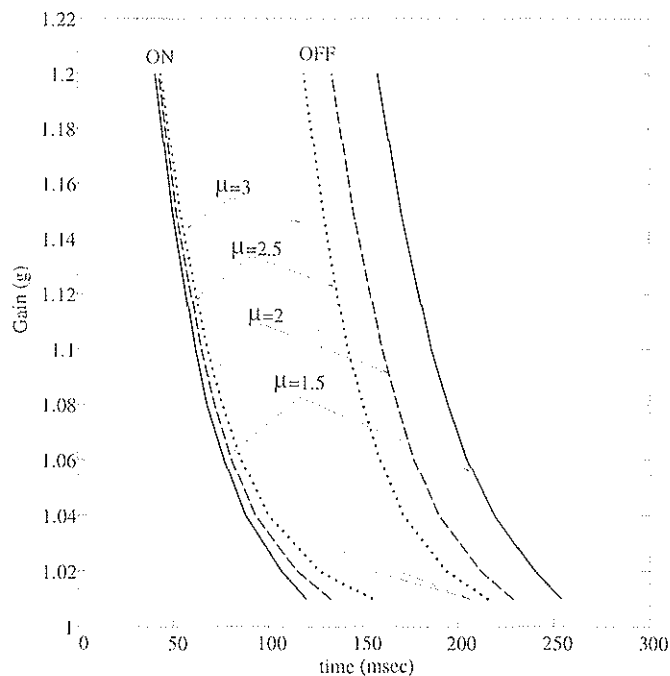


Figure 15. Gain g in (2) is plotted along the ordinate and resonance onset and offset times plotted along the abscissa. There is one pair of curves for each value of the habituation parameter μ with $\lambda = 0.1$. Other parameters are the same as Figure 10, except $\gamma = 0.097$ and $\theta = 0.22$.

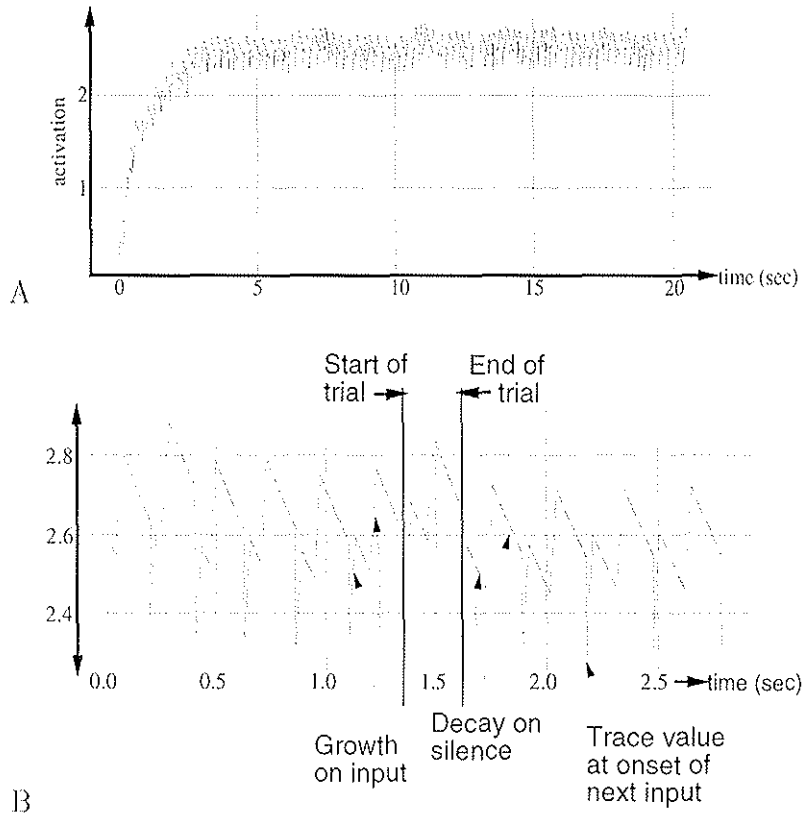


Figure 16. (A) Response of input rate estimator r in (9) through time during a simulation with 100 trials initialized by 10 warm-up trials. The main portion of inter-trial intervals are deleted (see text). (B) Detailed view of input rate estimator during one randomly selected trial for each of the silent intervals in the stimulus distribution. The inter-trial intervals are deleted as in (A).

silent intervals used in the trials increase, which is reflected in the increasing time for decay between intervals of growth.

10. Simulation of Repp (1980) Psychometric Functions

10.1 Method

The simulation procedure is directly analogous to the Repp (1980) experimental paradigm (Section 3). Each simulation consisted of a series of one hundred trials. For each trial, the network received two input pulses of fixed amplitude and duration, separated by a silent interval randomly selected from the appropriate Repp (1980) distribution (see Figure 3). For each presented silent interval, a count was made of the number of trials that yielded two output responses. Outputs were counted using the detection method described below. Dividing this count by the total number of trials at that silent interval gave the two-stop decision probability. The simulation was run eight times (once for each “subject”) for each of the six Repp (1980) distributions. For each silent interval, a count was made of the number of trials that yield two output responses, which was divided by the total number of trials to obtain the two-stop decision probability for that value of silent interval. Outputs are counted using the detection method described below. Simulations began with ten warm-up trials, as done by Repp (1980), which alternately used the shortest and longest silent interval between the two inputs. Stop clusters are represented by an input pulse to p_1 followed by an input pulse to p_2 . Geminate stops are represented by two input pulses to p_1 . The ISI

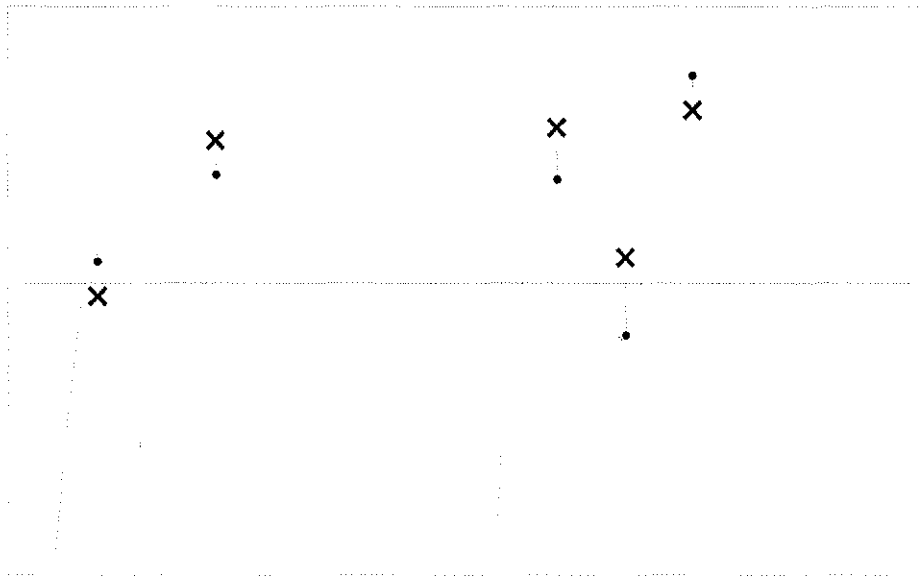


Figure 17. Typical list chunk activation traces plotted with the values used to decide if output is detected. (X) is a sample value after noise is added. In the examples shown, the first peak on the left, and the trough at right would not be detected, and so only one peak would be reported in each case.

of 2.5 seconds used by Repp (1980) can be deleted from the simulation trials because the network variables w_j , u_j , z_{jw} and z_{ju} reach their resting values within that period. Using the simulator event scheduler, the variables are set to these values – zero for activations w_j and u_j , unity for the transmitter levels z_{jw} and z_{ju} – at the start of each trial.

10.2 Detection of Output Responses

To simulate the Repp (1980) experiment, it is necessary to define the detection of a response, both its onset and offset. A simple definition for detecting output is that the list category activity u_j exceeds the resonance threshold, θ . Parameter θ is called a resonance threshold because list activation can exceed θ only after top-down feedback from the category node has begun. In principle, neural responses are noisy, so we seek a simple detection rule that looks not at the instantaneous threshold crossing, but reflects the likelihood that the response will exceed threshold before reset. The chosen strategy for detection is to wait for the peak list category response and compare that value to the detection threshold. Perceived silence exists during intervals in time between successive suprathreshold category responses. Thus, to detect geminate stops, category activation in response to the first phone must collapse below threshold before the onset of the second phone reinforces the response. Using the same likelihood principle as before, the strategy in this case is to wait for a trough in the response trajectory and compare that value to the detection threshold. The input to the decision processor is presumed to be perturbed by Gaussian distributed noise (Green and Swets, 1974).

Figure 17 demonstrates how phonetic percept detection was performed for the two cases of stop pairs. A decision was made whether the phone can be detected when the activity of the corresponding list node reaches a peak. At that time, a random noise sample is taken from a Gaussian distribution with zero mean and variance σ^2 and added to the peak response. The sum is compared to the fixed threshold, θ . After that time, a new peak in the activity of the same category node can be detected only if another decision is made that

the response has fallen below threshold, in accord with the proposition that the perception of silence between input phones corresponds to an interval of sub-threshold activation. The below-threshold decision is made when the output activation reaches a trough, using the same procedure as above. The peak height before adding noise may be regarded as the deterministic response of the system to inputs separated by a given silent interval. With noise, the peak corresponds to the mean value of a Gaussian distribution which has some calculable integral above the threshold criterion, which corresponds to the probability of exceeding threshold.

10.3 Simulation Results

Computer simulations of the model closely approximate the categorical decision boundaries reported by Repp (1980). These are shown in Figure 7. The simulated psychometric functions replicate the principal trends of the averaged subject data. Boundaries for all conditions shift with mean silent interval, indicating that the percept is approximately invariant with changes in mean silent interval. This shift property is a direct consequence of the automatic gain control in the model. Single-geminate boundaries are separated in time from the single-cluster boundaries by an interval approximately equal to the gap found in the data (≈ 150 msec, measured between the no-anchor curves for the two conditions). This gap is determined by the suprathreshold phase of the category node response. The single-cluster boundary occurs at the mean onset time of the response, where there is a 50% probability of detecting a suprathreshold response to the first input. The single-geminate boundary occurs at the mean offset time, where there is a 50% probability of detecting an interval of subthreshold activity prior to the second onset.

Boundary slopes for the geminate condition are also smaller than slopes for the cluster condition, as observed in the data. Not only the slope of the averaged boundary, but slopes for all individual geminate simulations are shallow compared to the cluster runs, indicating that the probability of detecting a subthreshold u value varies relatively slowly with silent interval when mean silent intervals are in the geminate regime. To understand this result, recall that by the detection procedure, probability bears a direct relation to the slope of the u trajectory within the time range of silent intervals being presented. Gain is low in the geminate regime, so u decays slowly with time in the temporal range of the silent intervals presented. Consequently, the probability of u being below the threshold at the end of the silent interval varies slowly with silent interval.

Figure 8 presents a more detailed comparison of the simulation and data. There are no individual subject data, although subject variability was reported to be high. Recent experimental results of Govindarajan and Cohen (1994) confirm the high subject variability in /ib/-/ga/ discriminations. Therefore, it is not possible to more completely characterize the degree of agreement between the model and the individual data. However, the mean values of the simulated percept probabilities were compared to the observed average probabilities. The error in modeling probability averaged over the six curves was 14%. The root mean square error was computed for each curve using $\sqrt{\sum_{j=1}^N (y_j - \bar{x}_j)^2 / N}$, where the y_j are the two-stop probabilities reported in Repp (1980) and the \bar{x}_j are the averaged simulated probabilities, j indexing silent interval.

The position of the simulated boundaries in the high anchor, single-cluster condition (case 1), and the low anchor, single-geminate condition (case 2) appear to be somewhat shifted with respect to the empirical boundaries, although, as indicated above, the error is unknown. If these differences are significant, they may suggest modifications to any or all of the factors which control u onset and offset times, namely, the gain function (2), the threshold θ , and the transmitter gates z_{ij} , as discussed above. Assuming the gain as a function of input rate has a monotone increasing slope, no simple change to the function will produce both a decrease in gain that delays onset times for case 1, and an increase in gain

that advances offset times for case 2, all else remaining unchanged. However, changes in the threshold θ can also effect shifts in the boundaries. If θ were not constant but increased with the duration of suprathreshold activity (Grossberg, 1976a; Rumelhart and Zipser, 1985), then thresholds under single-geminate input conditions would be higher than they are with a constant threshold, and, consequently, offset times would be earlier. It appears that combining this systematic change in threshold with a small change of parameters in (2) would improve the fit, but such refinements are unwarranted in the absence of additional data. The main purpose of the present simulations is to show that resonance concepts are sufficient to capture the main trends in the data through a real time simulation, and to encourage further experiments that are better designed to disclose details of resonant dynamics.

11. Mismatch Reset by an Orienting Subsystem

This section describes an alternative model mechanism for mismatch reset. Category reset as defined in Section 7 occurs when the category loses its bottom up support, coupled with the pressure of a relatively strong decay controlled by δ in equation (3). In Adaptive Resonance Theory (Carpenter and Grossberg, 1991, 1993; Grossberg, 1980), an orienting subsystem continuously monitors the degree of pattern matching between bottom-up input and top-down activation at the input stage. When the patterns are insufficiently matched, as reflected by some metric of the distance between the two pattern vectors, a novelty burst or “arousal wave” is triggered, which tends to inhibit active category nodes as it initiates a memory search for a better-fitting category, or hypothesis. As a result of this search, a new category, better matched to the input pattern, can become active.

An orienting subsystem was implemented for the model using the approach of Carpenter and Grossberg (1987b). Input and working memory activation vectors were compared using a normalized dot product rule, giving a matching value

$$m = \frac{\sqrt{\sum_i^N (I_i + w_i)^2}}{\|\vec{I}\| + \|\vec{w}\|} \quad (12)$$

where $\|\vec{x}\|$ is the Euclidean norm, $\sqrt{\sum_i^N x_i^2}$, of the vector $\vec{x} = (x_1, x_2, \dots, x_n)$. A nonspecific inhibitory arousal signal a is released as m falls below the *vigilance* level ρ . In particular,

$$a = \psi \max(0, \rho - m), \quad (13)$$

which is positive only when an input pattern is presented that conflicts sufficiently with \vec{w} to drive m below ρ . Subtracting the arousal signal a from the activity level (3) gives

$$\frac{du_j}{dt} = g(r) \{ (\beta - u_j) z_j w_j^+ - u_j (\delta + a) \}. \quad (14)$$

Once u_j is reset by a , working memory nodes that are associated with u_j lose top-down support and their activation decays passively, even without feed-forward inhibition, if they receive no feed-forward excitation. A stop cluster simulation using the reset mechanism defined above is shown in Figure 18. The reset signal is triggered by the onset of I_2 , which creates a mismatch at the working memory stage. The match m in (12) at the instant after I_2 appears is

$$m \approx \frac{\sqrt{I_2^2 + w_1^2}}{I_2 + w_1}, \quad (15)$$

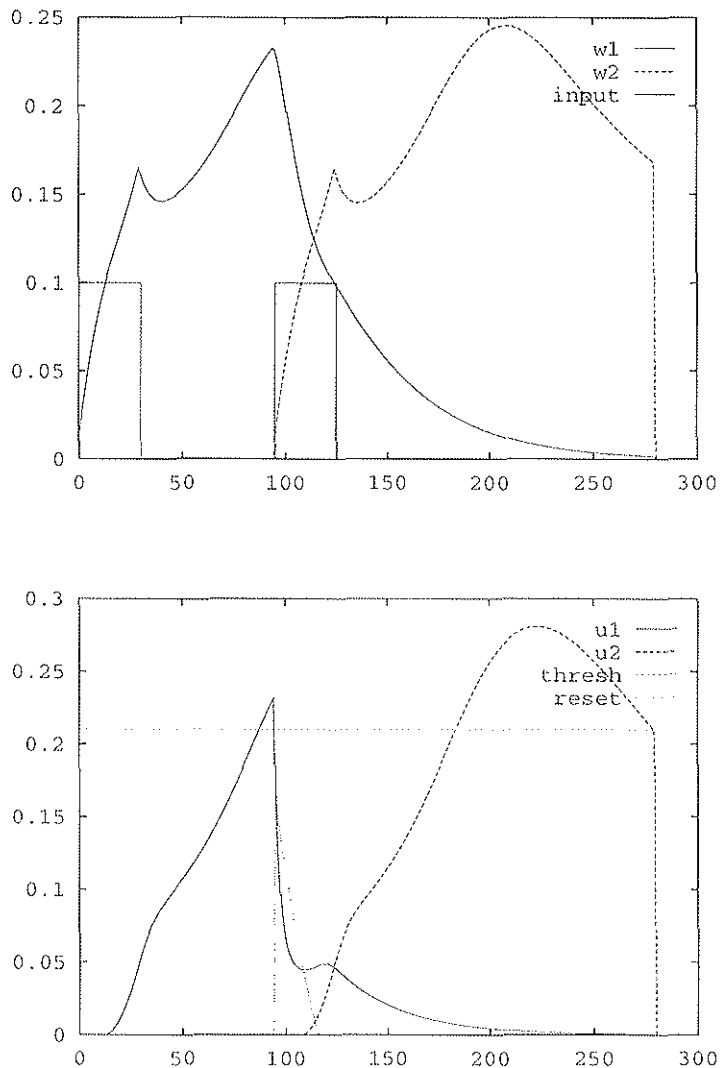


Figure 18. Activations w_j (top panel) and u_j (bottom panel) for a sequence of two phonemically unrelated inputs (rectangular pulses, top panel). Arousal signal given by equation (13) begins synchronously with second input (triangular burst, lower panel), and is plotted before scaling by coefficient $\psi = 20$, with vigilance $\rho = 0.9$. Other parameters are similar to those in Figure 10, except $\kappa = 0$ in equation (1).

which has a minimal value $1/\sqrt{2}$ for $I_2 = w_1$. A reset signal is sent if ρ exceeds that value, for some interval after the onset of I_2 , which is plotted as the triangular pulse in the lower graph of Figure 18. The reset inhibits active chunks in proportion to their activation, due to the shunting inhibitory term $-u_j a$ in (14), so u_1 drops sharply, quenching the resonance with w_1 .

12. Comparison with Alternative Models: FLMP, IAC, and TRACE

The ARTPHONE model is compared in this section with several other models to high-

light their similarities and differences.

The Fuzzy Logical Model of Perception (FLMP) of Massaro and colleagues (Massaro, 1989; Massaro and Cohen, 1991; Massaro and Oden, 1995) has had some notable successes simulating speech data. FLMP also shares some key features with the ART model, but differs from it in basic ways. In particular, the FLMP model's heuristics are closer to ART than is its computational instantiation.

Speaking heuristically, FLMP assumes that sensory systems activate bottom-up features that are matched against top-down prototypes. An identification decision is made using the relative goodness-of-match between these ingredients. Thus, as in ART, both bottom-up and top-down information are matched. To illustrate FLMP computations, suppose that prototypes R and L are used in a given task. Denote the i^{th} stimulus feature that supports R by f_i and the complementary feature that supports L by $1 - f_i$. Likewise, denote the j^{th} top-down context that supports R by c_j and its complement by $1 - c_j$. Then the degree of match to R and L are given by the products of bottom-up and top-down information; namely $R = f_i c_j$ and $L = (1 - f_i)(1 - c_j)$, and the probability of an R response is

$$P = \frac{f_i c_j}{f_i c_j + (1 - f_i)(1 - c_j)}. \quad (16)$$

Equation (16) shares some properties with ART. For example, it suggests that top-down context interacts with bottom-up signals, and that the match value is normalized against available alternatives. The ART matching rule and self-normalizing competitive dynamics also have these properties.

This being said, it needs to be noted that FLMP, computationally speaking, is an algebraic equation that is used to fit data through parameter estimation. There are no model internal representations. There is no emergent process from whose dynamics category boundaries can be estimated. Although Massaro and Cohen (1991) assume that a process such as f_i can take hold through time via the simple integration process

$$R_i = \frac{1}{2} e^{-\theta t} + f_i(1 - e^{-\theta t}), \quad (17)$$

this process is not linked to any system representation, and it proceeds at a constant rate which is insensitive to the external speech rate and to temporally nonuniform properties like mismatch reset or resonance. That is why FLMP is typically used to describe category boundaries using the alternatives R and L , rather than elapsed time t , as the independent variable.

Massaro and Cohen (1991) have argued that FLMP is more parsimonious and gives better data fits than models like TRACE (McClelland and Elman, 1986) and is thus preferable. Cutting *et al.* (1992) and Pitt (1995a, 1995b) have argued that FLMP's parsimony is of a type that allows it to fit data all too well, without regard to the fact that it may equally well fit data with different, even contradictory, processing implications. For example, to show how bottom-up information f_i from an initial speech segment and top-down context c_j from the following speech segment interact, Massaro and Oden (1995) simply compute $f_i c_j$. No analysis is given of how the system knows how to do this or why, for example, other combinations such as $f_i c_i$ are not also computed. More generally, the definitions of the feature f_i and context c_j are not given internal structure. For example, in the case of phonemic restoration, what are the "features" in the noise that precedes "eel" in "noise-eel"? Are they individual spectral components? If not, then how can they be multiplicatively matched to select only those spectral components that are consistent with the prototype? However, in applications of FLMP to the present, the features have not been spectral components. Likewise, it is not clear how FLMP could simulate nontrivial temporal properties of speech, such as the 150

msec shift of the /ib/-/ba/ category boundary in Figure 4 or the separate boundaries given different mean silent intervals. Without internal representations and temporal dynamics to constrain a model such as FLMP, it must remain a fundamentally incomplete model of cognitive processing, notwithstanding its proven ability to fit some speech data very well.

The interactive activation model (IAM) of letter perception (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) and the TRACE model of speech perception (McClelland and Elman, 1986) are closely related, both historically and conceptually, to ART models, but also exhibit some notable differences. As in Grossberg (1978a), IAM posited bottom-up and top-down influences on letter perception. Unlike ART, the 1981-82 version of IAM posited different connections and processing levels than did Grossberg (1978a). In particular, IAM assumed that both excitatory and inhibitory connections exist between levels, such that (say) compatible letters excite target words whereas incompatible letters inhibit target words. In addition, IAM posited separate phoneme, letter, and word levels of processing. In all ART models (e.g. Grossberg, 1976b, 1980a), all connections between levels (both bottom-up and top-down) are excitatory, and all inhibitory connections are contained within a level. For purposes of language-related processing, the levels represent items in working memory and list chunks (Grossberg, 1978a), rather than letters and words.

The ART processing levels and connections are preferable to those of IAM because the latter cannot stably learn their letter and word representations, and are incompatible with various data about letter recognition, as was pointed out in Grossberg (1984, 1987). Given how things later turned out, it is of some interest that these deficiencies of IAM were pointed out to the authors as early as 1980, half a year before the IAM articles were submitted for publication. In the letter Grossberg (1980b) to Jay McClelland, it was noted that the ART model in Grossberg (1978a) “(1) explains the boundary effects on word recognition as part of a theory of how temporal order information unfolds through time over item representations, (2) uses these patterned representations as a basis for code (or chunk) learning, (3) shows how subfields of chunks sensitive to different length lists mask each other using a principle of self-similarity as a basis for resolving uncertain data, (4) explains how the feedback templates from words to letters are learned and matched against letter codes, (5) shows why familiar letters need to have word-like representations to distinguish between representations that are reset by rehearsal and representations that are reset only by competition from other representations. At bottom, [ART] differs from [IAM] by showing how constraints on learning and code stability force the laws for competition too.” This communication was inspired by the fact that the letter and word recognition analyses in Grossberg (1978a) anticipated the data that IAM was used to model in a lecture that McClelland gave at MIT in 1980.

Several years later, the IAM postulates were replaced by ART postulates in McClelland (1985) and McClelland and Elman (1986). At this time, McClelland (1985) also recommended that IAM be viewed, not as a model, but as a framework in which one could avoid “worrying about the plausibility of assuming that they provide an adequate description of the actual implementation” (p. 144). This attitude was criticized in Grossberg (1987) because it would prevent falsifiability of a model. IAM could not be disproved because its postulates could be freely changed to those of other models, including earlier models such as ART.

In particular, the McClelland and Elman (1986) speech model incorporated some basic ART postulates: “units on different levels that are mutually consistent have mutually excitatory connections while units on the same level that are inconsistent have mutually inhibitory connections. The interactive activation model included inhibitory connections between [levels]...more recent versions...eliminate these between-level inhibitory connections, since these connections can interfere with successful use of partial information. This feature of TRACE plays a very important role in its ability to simulate a number of empirical phenomena” (McClelland and Elman, 1986, pp. 10-12). On the other hand, TRACE makes other assumptions that would make it very hard for it to explain the type of data discussed herein

and in earlier ART language analyses; e.g., Cohen and Grossberg (1986) and Grossberg (1978a, 1978b, 1986).

A key problem of TRACE is that “it requires massive duplication of units and connections, copying over and over again the connection patterns that determine which features activate which phonemes and which phonemes activate which words” (p. 77). This happens in TRACE because it posits that each event is recopied multiple times at every moment, or “time slice”, during which it occurs. The TRACE model thus does not operate in real time. It does not treat time as an independent variable. Rather, it treats time as a structural variable that is used to separate events into different time slices. As a result, TRACE cannot analyse variable-rate speech, as in the Repp (1980) data. It is also not possible for the model to understand generalization effects, as they are usually understood, because each word representation has multiple copies on multiple time slices. Indeed “there is a unit for every word in every time slice” (p. 18). Learning is also rendered difficult for the same reason. In particular, how would learning of a representation on one time slice influence a representation of the same information on another time slice?

The dynamical equations used by IAM and TRACE are related to the shunting equations used in ART, as McClelland and Rumelhart (1981) and McClelland and Elman (1986) both note. As in the shunting equation (1), the TRACE dynamical equations describe activations that remain bounded due to multiplication by shunting terms. However, the TRACE equations were modified so that they do not have a plausible neural interpretation and lose the main property of the shunting on-center off-surround networks; namely, the ART equations compute ratios of their inputs and do not saturate when the total input becomes large. These properties are crucial for designing working memories that can explain temporal order data and whose activation patterns can be stably learned by list chunks (Bradski, Carpenter, and Grossberg, 1992, 1994, Grossberg, 1973, 1978a, 1978b). TRACE loses these properties because the sum $\Sigma = E - I$ of excitatory inputs E and inhibitory inputs $-I$ to a cell with activity x is multiplied in TRACE by a shunting term $(\beta - x)$ if Σ is positive, and by a shunting term $-(x + \gamma)$ if Σ is negative. This discrete switch between terms $(\beta - x)$ and $-(x + \gamma)$ has no obvious physical interpretation. Due to this switching term, x remains bounded between the values β and $-\gamma$, but becomes insensitive to E and I as Σ becomes large in absolute value. In contrast, w_j in (1) does not lose its sensitivity to input ratios as the total input becomes large. This is because the excitatory on-center inputs in (1) multiply the shunting term $(\beta - w_j)$ while the inhibitory off-surround terms simultaneously multiply the shunting term $-w_j$. Only then are the two terms summed. In summary, adding E and $-I$ inputs before shunting them leads to qualitatively different properties than shunting them individually before adding them.

The TRACE model omits two of the most important ART mechanisms for building a more complete theory of speech and language processing. In particular, TRACE does not develop mechanisms of resonance and reset. The authors claim that “it keeps straight what occurred when in the speech stream” (p. 75), but it does this only at the price of replicating all events and their representations in multiple time slices. Despite this maneuver, the model cannot explain significant backward effects in time over silent intervals. Indeed, silence is itself treated as feature that is input to the network into a new time slice at each moment when silence occurs. Such a rigidly clocked silence cannot naturally explain the 150 msec shift in the category boundaries described in Repp (1980) and simulated herein. In contrast, a concept of resonance allows fusion events, as in the [iba] percept of Figure 4, to span an unusually long silence interval, and allows future data, such as [g] in the percept [iga] to influence past input activations before they reach resonance.

The two types of ART reset – category collapse and mismatch reset – also do not occur in TRACE. These reset mechanisms clarify how resonances can be terminated despite the positive feedback that occurs due to bottom-up and top-down excitatory signals. In contrast, TRACE has no natural real-time mechanisms for resetting representations. Instead, it treats

silence as a structural feature which can be used to inhibit non-silent representations via lateral inhibition from silence nodes to other nodal representations.

TRACE also does not implement an ART matching rule. For example, "If higher levels insist that a particular phoneme is present, then the unit for that phoneme can be activated ...; then the learning mechanism can 'retune' the detector..." (p. 75). This property implies that learning in the TRACE model, were it ever implemented, would be unstable through time (Carpenter and Grossberg, 1987a; Grossberg, 1980). This way of using top-down feedback also implies that TRACE cannot explain phonemic restoration data, as in Section 1, in which silence remains silent after top-down feedback acts, and a reduced set of spectral components in a noise input leads to a correspondingly degraded consonant sound.

The authors of TRACE are aware of some of these difficulties. They end their article by noting that a "fundamental deficiency of TRACE is that it requires massive duplication of units. However, it remains necessary to keep straight the relative temporal location of different...activations...we need to have it both ways...so that we can continue to accommodate both left and right contextual effects" (p. 77). TRACE uses some basic ART postulates to partially accomplish this. On the other hand, by not incorporating a true resonance event, and all that goes with it, the TRACE model loses the ability to operate in real time and to self-organize. We suggest, as in Grossberg (1978a, 1986) that many of these problems vanish when a speech percept is analysed as a resonant wave. Then silence can be interpreted as a temporal discontinuity in the resonant wave, rather than as a built-in silence feature in a hard-wired series of time slices.

13. Discussion: How General are Resonant Dynamics in the Brain?

This article describes and simulates the ARTPHONE neural network model for rate-invariant phonetic perception that quantitatively develops aspects of the speech and word recognition model introduced in Grossberg (1978a, 1986); also see Cohen and Grossberg (1986) and Grossberg and Stone (1986). The ARTPHONE model uses list chunking nodes that categorize sequences of phonetic items while supporting the storage of consistent items in working memory using top-down feedback. Feedback to working memory nodes, in turn, increases bottom-up input to the associated list nodes, leading to a resonant response identified with the speech percept.

The long-lasting resonance time scale provides an explanation of why the geminate categorical curves in Figure 4 are shifted 150 msec beyond those of the cluster curves. The collapse of resonant responses due to habituation helps to account for the finite span of conscious percepts in general and the single-geminate boundary in particular. An earlier version of the model, briefly reported in Boardman, Cohen, and Grossberg (1993), could not explain the 150 msec shift of the categorical boundary in the geminate case because it did not incorporate ART resonant matching.

In the cluster case, when new inputs, such as /g/, are inconsistent with an active expectation, such as that of /b/, as reflected in the pattern of top-down feedback, then active list chunks are rapidly reset, thereby explaining how the resonance time scale is actively cut short in the cluster case. Network activation rates are also globally modulated based on a running average of input signal density so that the time required to segregate or integrate responses scales with mean input presentation rate. These properties allow computer simulations of the model to closely approximate human subject performance in psychophysical discriminations of stop-consonant pairs.

The model simulation will be further extended in the future to include multiple scales for list nodes. List nodes for list sizes greater than one are, for example, needed to address the "gray chip"/"great ship" dichotomy reported by Repp *et al.* (1978). The list nodes would then, as suggested in Grossberg (1986), incorporate multiple scale chunking network properties, also called masking field properties, that were referred to in Section 2. Using such an enhanced network, a list category node for /gret/ could mask the nodes for /gre/ under

conditions that support activation of the /t/ item. It would then be possible to investigate the role of segment duration and speaking rate on grouping dynamics for several list lengths.

The central message of the present simulations is to illustrate how a model wherein conscious speech is an emergent property of a resonant process can be used to quantitatively explain difficult psychometric data about variable-rate speech categorization. Such data exemplify what Bregman (1990) has called the *schema-based* segregation process, to distinguish it from the *primitive* streaming process whereby multiple acoustic sources can be segregated from one another using cues such as pitch and location, as in the cocktail party problem. Primitive streaming data have recently been modeled using a neural architecture, called the ARTSTREAM model, in which ART resonant dynamics again obtain, here between multiple spectral representations and pitch representations of acoustic data (Govindarajan *et al.*, 1994).

In the ARTSTREAM model, the incoming auditory signal gets preprocessed by the ears' mechanical and neurophysiological filters, which divide sounds into groups of similar frequencies. The spectral, or frequency, components of a sound stream serve as inputs for multiple spectral stream layers, which each convert the incoming signal into a spatial map of frequencies. As a result, a specific sound activates a specific spatial pattern of activation across the spectral stream cells of all model streams. This representation is analogous to the working memory of the ARTPHONE model.

Each spectral stream layer emits bottom-up signals to its pitch stream layer, which plays the role of the category layer in the ARTPHONE model. Between layers, the bottom-up pathways act like a type of harmonic sieve that filters the spectrum so that only certain harmonically related frequencies can activate a pitch node within each pitch stream. The filtered bottom-up signals activate multiple representations of a sound's pitch across the several streams at the pitch stream level. These pitch representations compete to select a single winning pitch node, which becomes active much as at the list chunking layer of the ARTPHONE model.

The winning pitch node inhibits the redundant pitch representations in other pitch streams, as it sends top-down matching signals back to its spectral stream level. These top-down signals realize the ART matching rule by exciting spectral nodes whose harmonically related frequencies are consistent with the selected pitch, and inhibiting all other frequencies within its stream. As a result only those spectral nodes that receive simultaneous bottom-up and top-down signals can remain active within that stream. This leads to a spectral-pitch resonance within the stream of the winning pitch node.

This resonance binds together the frequency components that correspond to an auditory source with a prescribed pitch. All the suppressed frequency components in this stream are then freed to activate other spectral streams and to resonate with a different pitch node in a different pitch stream. The net result is multiple spectral-pitch resonances, each selectively grouping together the frequencies that correspond to a distinct auditory source. The model shows how a given stream can track changes through time in each source's pitch in a manner that simulates psychophysical data.

In summary, both the ARTSTREAM and ARTPHONE models employ similar resonant dynamics that are specialized to deal with the different invariant properties of the inputs that they process. It therefore appears that resonant matching processes may play a role at multiple levels of auditory and speech processing to construct coherent representations of acoustic objects from the jumble of noise and harmonics that relentlessly bombards our ears throughout life.

How generally do similar resonant dynamics occur in other brain systems? To approach this question, it is important to realize that ART dynamics have been proposed to solve the general stability-plasticity dilemma of how the brain can rapidly learn new information without being forced into catastrophically forgetting previously learned information (Grossberg, 1980). This hypothesis raises the question of whether similar ART principles and mecha-

October 20, 1995

nisms are used to enable other brain systems than the auditory system to adapt to their changing input environments, perhaps with specialized properties that have evolved to cope with the different invariant properties of the inputs experienced by these systems. Grossberg (1995) has reviewed evidence that, indeed, ART mechanisms of attentive top-down matching and resonance are also employed in brain systems for early vision, visual object recognition, somatosensory recognition, and adaptive sensory-motor control. The ability of ART systems to rapidly and stably learn in real time to recognize large amounts of information in response to a rapidly changing environment has also led to their use in a wide variety of technological applications, ranging from airplane design and medical data base prediction to remote sensing and the control of mobile robots; see Carpenter and Grossberg (1994, 1995) for some references.

An exciting prospect for future research is to develop a precise understanding of how the shared organization principles in all the brain systems that undergo recognition learning, categorization, and prediction have been specialized through evolution, development, and learning for processing their own types of data. No less exciting is the prospect that the existence of such similar dynamics across modalities, and across levels of processing within modalities, promise to clarify how the brain integrates multiple sources of information into unified moments of conscious experience.

References

- Baddeley, A.D. (1986). **Working Memory**, Oxford Psychology Series No. 11. New York, NY: Oxford University Press.
- Bashford, J.A. Jr. and Warren, R.M. (1987). Multiple phonemic restorations follow the rules for auditory induction. *Perception and Psychophysics*, **42**, 114-121.
- Boardman, I. and Bullock, D. (1991). A neural network model of serial order recall from short-term memory. In **Proceedings of the international joint conference on neural networks**, Seattle, 1991, **II**, 879-884. Piscataway, NJ: IEEE Service Center.
- Boardman, I., Cohen, M. A., and Grossberg, S. (1993). Variable rate working memories for phonetic categorization and invariant speech perception. **Proceedings of the World Conference on Neural Networks (WCNN-93)**. Hillsdale, NJ: Lawrence Erlbaum Associates, **III**, 2-5.
- Bradski, G., Carpenter, G. A., and Grossberg, S. (1992). Working memory networks for learning temporal order with application to three-dimensional visual object recognition. *Neural Computation*, **4**, 270-286.
- Bradski, G., Carpenter, G.A., and Grossberg, S. (1994). STORE working memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics*, **71**, 469-480.
- Bregman, A.S. (1990). **Auditory scene analysis: The perceptual organization of sound**. Cambridge, MA: MIT Press.
- Carpenter, G.A. and Grossberg, G. (1983). A neural theory of circadian rhythms: The gated pacemaker. *Biological Cybernetics*, **48**, 35-59.
- Carpenter, G.A. and Grossberg, S. (1984). A neural theory of circadian rhythms: Aschoff's rule in diurnal and nocturnal mammals. *American Journal of Physiology (Regulatory, Integrative, and Comparative Physiology)*, **247**, R1067-R1082.
- Carpenter, G.A. and Grossberg, S. (1985). A neural theory of circadian rhythms: Split rhythms, after-effects, and motivational interactions. *Journal of Theoretical Biology*, **113**, 163-223.
- Carpenter, G.A. and Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, **37**, 54-115.
- Carpenter, G.A. and Grossberg, S. (1987b). ART2: Stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics*, **26**, 4919-4930.
- Carpenter, G.A. and Grossberg, S. (1990). ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, **3**, 129-152.
- Carpenter, G.A. and Grossberg, S. (Eds.) (1991). **Pattern Recognition by Self-Organizing Neural Networks**. Cambridge, MA: MIT Press.
- Carpenter, G.A. and Grossberg, S. (1992). A self-organizing neural network for supervised learning, recognition, and prediction. *IEEE Communications Magazine*, **30**, 38-49.
- Carpenter, G.A. and Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences*, **16**, 131-137.
- Carpenter, G.A. and Grossberg, S. (1994). Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction. In V. Honavar and L. Uhr (Eds.), **Artificial intelligence and neural networks: Steps towards principled predictions**. San Diego: Academic Press, pp. 387-421.
- Carpenter, G.A. and Grossberg, S. (1995). A neural network architecture for autonomous learning, recognition, and prediction in a nonstationary world. In S.F. Zornetzer, J.L. Davis, C. Lau, and T. McKenna (Eds.) **An introduction to neural and electronic networks**, 2nd edition. San Diego, CA: Academic Press.

- Cohen, M.A. and Grossberg, S. (1986). Neural dynamics of speech and language coding: developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, **5**, 1-22.
- Cohen, M.A. and Grossberg, S. (1987). Masking fields: A massively parallel architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, **26**, 1866-1891.
- Cohen, M.A., Grossberg, S., and Stork, D.G. (1988). Speech perception and production by a self-organizing neural network. In Y.C. Lee (Ed.), **Evolution, Learning, Cognition, and Advanced Architectures**. Hong Kong: World Scientific Publishers.
- Dorman, M.F., Raphael, L.J., and Lieberman, A.M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, **65**, 1518-1532.
- Cutting, J.E., Bruno, N., Brady, N.P., and Moore, C. (1992). Selectivity, scope, and simplicity of models; A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.
- Francis, G. and Grossberg, S. (1995a). Cortical dynamics of boundary segmentation and reset: Persistence, afterimages, and residual traces. *Perception*, in press. Tech Report CAS/CNS-TR-95-002, Boston, MA: Boston University
- Francis, G. and Grossberg, S. (1995b). Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours. *Vision Research*, in press. Tech Report CAS/CNS-TR-94-011, Boston, MA: Boston University
- Francis, G., Grossberg, S., and Mingolla, E. (1994). Cortical dynamics of feature binding and reset: Control of visual persistence. *Vision Research*, **34**, 1089-1104.
- Gaudio, P. and Grossberg, S. (1991). Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, **4**, 493-504.
- Govindarajan, K.K. and Cohen, M.A. (1994). Influence of silence duration distribution in perception of stop consonant clusters, *Journal of the Acoustical Society of America*, **95**, 2978.
- Govindarajan, K.K., Grossberg, S., Wyse, L., and Cohen, M.A. (1994). A neural network model of auditory scene analysis and source segregation. Technical Report CAS/CNS-TR-94-039. Boston, MA: Boston University.
- Green, D. M. and Swets, J. A. (1974). **Signal Detection Theory and Psychophysics**. New York, NY: Kreiger Press.
- Grossberg, S. (1968). Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences*, **60**, 758-765.
- Grossberg, S. (1969). On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology*, **22**, 325-364.
- Grossberg, S. (1972). A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences*, **15**, 253-285.
- Grossberg, S. (1973). Contour enhancement, short term memory and constancies in reverberating networks. *Studies in Applied Mathematics*, **52**, 217-257. Reprinted in S. Grossberg (Ed.) (1982). **Studies of Mind and Brain**, Boston, MA: Reidel Press.
- Grossberg, S. (1976a). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**, 121-234.
- Grossberg, S. (1976b). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions, *Biological Cybernetics*. **23**, 187-202.
- Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps and plans. In R. Rosen and F. Snell (Eds.), **Progress in**

- Theoretical Biology, Vol. 5.** Academic Press, New York, 233–375. Reprinted in S. Grossberg (Ed.) (1982). **Studies of Mind and Brain.**, Boston, MA: Reidel Press.
- Grossberg, S. (1978b). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, **3**, 199–219.
- Grossberg, S. (1980a). How does the brain build a cognitive code? *Psychological Review*, **1**, 1–51.
- Grossberg, S. (1980b). Letter to Jay McClelland, September 15, 1980.
- Grossberg, S., (Ed.) (1982). **Studies of Mind and Brain.** Boston, MA: Reidel Press.
- Grossberg, S. (1984). Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory*, **7**, 263–283.
- Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Pattern Recognition by Humans and Machines, Vol. 1: Speech Perception.** New York, NY: Academic Press.
- Grossberg, S. (1987). Competitive learning: From interactive activations to adaptive resonance. *Cognitive Science*, **11**, 23–63.
- Grossberg, S. and Grunewald, A. (1995). Temporal dynamics of binocular disparity processing with corticogeniculate interactions. Technical Report CAS/CNS-TR-94-025, Boston, MA: Boston University
- Grossberg, S. and Kuperstein, M. (1986). **Neural Dynamics of Adaptive Sensory-Motor Control: Ballistic Eye Movements**, Amsterdam, NETH: Elsevier Publisher. Expanded edition, 1989, Elmsford, NY: Pergamon Press.
- Grossberg, S. (1995). The attentive brain. *American Scientist*, **83**, 438–449.
- Grossberg, S. and Stone, G.O. (1986a). Neural dynamics of attention switching and temporal order information in short-term memory. *Memory and Cognition*, **14**, 451–468.
- Grossberg, S. and Stone, G.O. (1986b). Neural dynamics of word recognition and recall: Attentional priming, learning and resonance. *Psychological Review*, **93**, 46–74.
- Irvine, D.R.F. (1986). **Progress in sensory physiology 7.** Berlin, DE: Springer-Verlag.
- Massaro, D.W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, **21**, 398–421.
- Massaro, D.W. and Cohen, M.M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology*, **23**, 558–614.
- Massaro, D.W. and Cohen, M.M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, **122**, 115–124.
- Massaro, D.W. and Oden, G.C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1053–1064.
- McClelland, J.L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science*, **9**, 113–146.
- McClelland, J.L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375–407.
- McClelland, J.L. and Rumelhart, D.E. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1–86.
- Miller, G.A. (1956). The magic number seven plus or minus two. *Psychological Review*, **63**, 81–97.

- Parker, D.B. (1982). Learning-logic. Invention Report, 581-64. File 1, Office of Technology Licensing. Stanford University, October 1982.
- Pickett, J.M. and Decker, L.R. (1960). Time factors in perception of a double consonant. *Language and Speech*, **3**, 11-17.
- Pickles, J.O. (1988). **An Introduction to the Physiology of Hearing**, 2nd edition. San Diego, CA: Academic Press.
- Pitt, M.A. (1995a). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1037-1052.
- Pitt, M.A. (1995b). Data fitting and detection theory: A reply to Massaro and Oden (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1065-1067.
- Repp, B.H. (1980). A range-frequency effect on perception of silence in speech. Haskins Laboratories Status Report on Speech Research, **SR-61**, 151-165.
- Repp, B.H. (1988). Integration and segregation in speech perception. *Language and Speech*, **31**, 239-271.
- Repp, B. H., Liberman, A.M., Eccardt, T., and Pesetsky, D. (1978). Perceptual integration of temporal cues for stop, fricative, and affricative manner. *Journal of Experimental Psychology: Human Perception and Performance*, **4**, 621-637.
- Rosetti, R. (1994). Gemination of Italian stops. *Journal of the Acoustical Society of America*, Program: 127th Meeting of the ASA, **95**(5)2, 2874.
- Rumelhart D.E., Hinton, G.E. and Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), **Parallel distributed processing**. Cambridge, MA: MIT Press.
- Rumelhart, D.E. and McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60-94.
- Rumelhart, D.E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75-112.
- Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology, *Journal of Experimental Psychology: General*, **110**, 474-494.
- Samuel, A.G., van Santen, J.P.H, and Johnston, J.C. (1982). Length effects in word perception: we is better than I but worse than you or them. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 91-105.
- Schvaneveldt, R.W. and McDonald, J.E. (1981). Semantic context and the encoding of words: Evidence for two modes of stimulus analysis. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 673-687.
- Warren, R.M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, **96**, 371-383.
- Warren, R.M. and Sherman, G.L. (1974). Phonemic restorations based on subsequent context. *Perception and Psychophysics*, **16**, 150-156.
- Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. Thesis, Harvard University, Cambridge, Massachusetts.