The Adaptive Self-organization of Serial Order in Behavior: Speech, Language, and Motor Control*

Stephen Grossberg

Department of Mathematics, Boston University,
Boston, Massachusetts 02215

I. INTRODUCTION: PRINCIPLES OF SELF-ORGANIZATION IN MODELS OF SERIAL ORDER: PERFORMANCE MODELS VERSUS SELF-ORGANIZING MODELS

The problem of serial order in behavior is one of the most difficult and far-reaching problems in psychology (Lashley, 1951). Speech and language, skilled motor control, and goal-oriented behavior generally are all instances of this profound issue. This chapter describes principles and mechanisms that have been used to unify a variety of data and models, as well as to generate new predictions concerning the problem of serial order.

The present approach differs from many alternative contemporary approaches by deriving its conclusions from concepts concerning the adaptive self-organization (e.g., the development, chunking, and learning) of serial behavior in response to environmental pressures. Most other approaches to the problem, notably the familiar information processing and artificial intelligence approaches, use performance models for which questions of self-organization are raised peripherally, if at all. Some models discuss adaptive issues but do not consider them in a real-time context. A homunculus is often used, either implicitly or explicitly, to make the model work. Where a homunculus is not employed, models are often tested numerically in such an impoverished learning environment

* Supported in part by National Science Foundation Grants NSF IST-80-00257 and NSF-IST-8417756 and Office of Naval Research Contract ONT-N00014-83-K0337.

that their instability in a more realistic environment is not noticed. These limitations in modeling approaches have given rise to unnecessary internal paradoxes and predictive limitations within the modeling literature. I suggest that such difficulties are due to the facts that principles and laws of self-organization are rate-limiting in determining the design of neural processes, and that problems of self-organization are the core issues that distinguish psychology from other natural sciences, such as traditional physics.

In light of these assertions, it is perhaps more understandable why a change of terminology or usage of the same concepts and mechanisms to discuss a new experiment can be hailed as a new model. The shared self-organizing principles that bind the ideas in one model to the ideas in other models are frequently not recognized. This style of model building tends to perpetuate the fragmentation of the psychological community into non-interacting specialties rather than foster the unifying impact whereby modeling has transformed other fields.

The burgeoning literature on network and activation models in psychology has, for example, routinely introduced as new ideas concepts that were previously developed to explain psychological phenomena in the neural modeling literature. Such concepts as unitized nodes, the priming of short-term memory, probes of long-term memory, automatic processing, spreading activation, distinctiveness, lateral inhibition, hierarchical cascades, and feedback were all quantitatively used in the neural modeling literature before being used by experimental psychologists. Moreover, the later users have often ignored the hard-won lessons to be found in the neural modeling literature.

The next section illustrates some characteristic difficulties of models and how they can be overcome by the present approach (this discussion can be skipped on a first reading).

II. MODELS OF LATERAL INHIBITION, TEMPORAL ORDER, LETTER RECOGNITION, SPREADING ACTIVATION, ASSOCIATIVE LEARNING, CATEGORICAL PERCEPTION, AND MEMORY SEARCH: SOME PROBLEM AREAS

II.A. Lateral Inhibition and the Suffix Effect

From a mathematical perspective, a model that uses lateral inhibition is a competitive dynamical system (Grossberg, 1980a). Smale (1976) has proven that the class of competitive dynamical systems contains systems capable of exhibiting arbitrary dynamical behavior. Thus to merely say that lateral inhibition is at work is, in a literal mathematical sense, a vacuous statement. One needs to define precisely the dynamics that one has in mind before anything of scientific value can be gleaned. Even going so far as to say that the inhibitory feedback between nearby populations is linear says nothing of interest, because linear feedback can cause such varied phenomena as oscillations that never die out or the persistent storage of short-term memory patterns, depending on the anatomy of the network as a whole (Cohen & Grossberg, 1983; Grossberg, 1978c, 1980a).

An imprecise definition of inhibitory dynamics will therefore inevitably produce unnecessary controversies, as has already occurred. For example, Crowder's (1978) explanation of the suffix effect (Dallett, 1965) and Watkins and Watkins's (1982) critique of the Crowder theory both focus on the purported property of recurrent lateral inhibition that an extra suffix should weaken the suffix effect due to disinhibition. However, this claim does not necessarily hold in certain shunting models of recurrent lateral inhibition that are compatible with the suffix effect (Grossberg, 1978a, 1978e). This controversy concerning the relevance of lateral inhibition to the suffix effect cannot be decided until the models of lateral inhibition used to analyze that effect are determined with complete mathematical precision. A type of lateral inhibition that avoids the controversy is derived from a rule of self-organization that guarantees the stable transfer of temporal order information from short-term memory to long-term memory as new items continually perturb a network (Section XXXIII).

II.B. Temporal Order Information in Long-Term Memory

A more subtle problem arises in Estes's (1972) influential model of temporal order information in long-term memory. Estes (1972, p. 183) writes: "The inhibitory tendencies which are required to properly shape the response output become established in memory and account for the long term preservation of order information." Estes goes on to say that inhibitory connections form from the representations of earlier items in the list to the representations of later items. Consequently, earlier items will be less inhibited than later items on recall trials and will therefore be performed earlier. Despite the apparent plausibility of this idea, a serious problem emerges when one writes down dynamical equations for how these inhibitory interactions might be learned in real time. One then discovers that learning by this mechanism is unstable because, as Estes realized, the joint activation of two successive network nodes is needed for the network to know which inhibitory pathway should be strengthened. As such an inhibitory pathway is strengthened, it can more strongly inhibit its receptive node, which is the main idea of the Estes model.

However, when this inhibitory action inhibits the receptive node, it undermines the joint excitation that is needed to learn and remember the strong inhibitory connection. The inhibitory connection then weakens, the receptive node is disinhibited, and the learning process is initiated anew. An unstable cycle of learning and forgetting order information is thus elicited through time. Notwithstanding the heuristic appeal of Estes's mechanism, it cannot be correct in its present form. All conclusions that use this mechanism therefore need revision, such as Rumelhart and Norman's (1982) discussion of typing and MacKay's (1982) discussion of syntax.

One might try to escape the instability problem that arises in Estes's (1972) theory of temporal order information by claiming that inhibitory connections are prewired into a sequential buffer and that many different lists can be performed from this buffer. Unfortunately, traditional buffer concepts (e.g., Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981) face design problems that are as serious as the instability criticism (Grossberg, 1978a). In this way, the important design problem of how to represent temporal order information in short-term and long-term memory without using either a traditional buffer or conditioned inhibitory connections is vividly raised. Solutions of these problems are suggested in Sections XII–XIX and XXXIV.

Letter and Word Recognition

A similar instability problem occurs in the work of letter perception of Rumelhart and McClelland (1982). They write: "Each letter node is assumed to activate all of those word nodes consistent with it and inhibit all other word nodes. Each active word node competes with all other word nodes . . . " (Rumelhart & McClelland, 1982, p. 61). Obviously, the selective connections between letter nodes and word nodes are not prewired into such a network at birth. Otherwise, all possible letter-word connections for all languages would exist in every mind, which is absurd. Some of these connections are therefore learned. If the inhibitory connections are learned, then the model faces the same instability criticism that was applied to Estes's (1972) model. Grossberg (1984b) shows, in addition, that if the excitatory connections are learned, then learning cannot get started.

The connections hypothesized by Rumelhart and McClelland also face another type of challenge from a self-organization critique. How does the network learn the difference between a letter and a word? Indeed, some letters are words, and both letters and words are pronounced using a temporal series of motor commands. Thus many properties of letters and words are functionally equivalent. Why, then, should each word compete with all other words, whereas no letter competes with all other words? An alternative approach is suggested in Section XXXVII, where it is suggested that the levels used in the Rumelhart and McClelland model are insufficient.

The McClelland and Rumelhart model faces such difficulties because it considers only performance issues concerning the processing of four-letter words. In contrast, the present approach considers learning and performance issues concerning the processing of words of any length. Its analysis of how a letter stream of arbitrary length is organized during real-time presentation leads to a process that predicts, among other properties, a word-length effect in word superiority studies (Grossberg, 1978e, Sect. 41; reprinted in Grossberg, 1982d, p. 595). Subsequent data have supported this prediction (e.g., Matthei, 1983; Samuel, van Santen, & Johnston, 1982, 1983). No such prediction could be made using Rumelhart and McClelland's (1982) model, since it is defined only for four-letter words. Moreover, the theoretical ideas leading to predictions such as the word-length effect are derived from an analysis of how letter and word representations are learned. An analysis of performance issues per se provides insufficient constraints on processing design.

II.D. Spreading Activation

Similar difficulties arise from some usages of ideas like spreading activation in network memory models. In Anderson (1976) and Collins and Loftus (1975), the amount of activation arriving at a network node is a decreasing function of the number of links the activation has traversed, and the time for activation to spread is significant (about 50–100 ms per link). By contrast, there is overwhelming neural evidence of activations that do not pass passively through nerve cells and that are not carried slowly and decrementally across nerve pathways (Eccles, 1952; Kuffler & Nicholls, 1976; Stevens, 1966). Rather, activation often cannot be triggered at nerve cells unless proper combinations of input signals are received, and when a signal is elicited it is carried rapidly and nondecrementally along nerve pathways. Although these ideas have been used in many neural network analyses of psychological data, their unfamiliarity to many psychologists is still a source of unnecessary controversy (Ratcliff & McKoon, 1981). Most spreading activation models are weakened by their insufficient concern for which nodes have a physical existence and which dynamical transactions occur within and between nodes. Both of these issues are special cases of the general question of how a node can be self-organized through experience.

Anderson's (1983) concept of a "fan effect" in spreading activation illustrates these difficulties. Anderson proposes that if more pathways lead away from a concept node, each pathway can carry less activation. In this view, activation behaves like a conserved fluid that flows through pipelike pathways. Hence, the activation of more pathways will slow reaction time, other things being equal. The number of pathways to which a concept node leads, however, is a learned property of a self-organizing network. The pathways that are strengthened by learning are a subset of all the pathways that lead away from the concept node. At the concept node itself, no evidence is available to label which of these pathways was strengthened by learning (Section III). The knowledge of which pathways are learned is only available by testing how effectively the learned signals can activate their recipient nodes. It is not possible, in principle, to make this decision at the activating node itself.

Since many nodes may be activated by signals from a single node, the network decides which nodes will control observable behavior by restricting the number of activated nodes. Inhibitory interactions among the nodes help to accomplish this task. Inhibitory interactions are not used in Anderson's (1983) theory, although it is known that purely excitatory feedback networks are unstable unless artificially narrow choices of parameters are made. Without postulating that activation behaves like a conserved fluid, a combination of thresholds and inhibitory interactions can generate a slowing of reaction time as the number of activated pathways is increased. In fact, the transition from a fan concept (associative normalization) to inhibitory interactions and thresholds was explicitly carried out and applied to the study of reaction time (Grossberg, 1968b, 1969c). This theoretical step gradually led to the realization that inhibitory interactions cause limited capacity properties as a manifestation of a fundamental principle of network design (Section XIX). Anderson (1983) intuitively justifies his fan concept in terms of a limited capacity for spreading activation, but he does not relate the limited capacity property to inhibitory processes.

II.E. Associative Learning and Categorical Perception

In the literature on associative learning, confusion has arisen due to an insufficient comparative analysis of the adaptive models that are available. For example, some authors erroneously claim that all modern associative models use "Hebbian synapses" (Anderson, Silverstein, Ritz, & Jones, 1977) and thus go on to equate important differences in processing capabilities that exist among different associative models. For example, in their discussion of long-term memory, Anderson et al. (1977) claim that the change in synaptic weight z_{ij} from a node v_i to a node v_j equals the product of the activity f_i of v_i with the activity g_j of v_j, where f_i and g_j may be positive or negative. If both f_i and g_j are negative, two inhibited nodes can generate a positive increment in memory, which is neurally unprecedented. Also, if f_i is positive and g_j is negative, a negative memory trace z_{ij} can occur. Later, if f_i is negative, its interaction with negative memory z_{ij} causes a positive activation of g_j. Thus an inhibited node v_i can, via a negative memory trace z_{ij}, excite a node v_j. This property is also neurally unprecedented. Both of these properties follow from the desire of Anderson et al. (1977) to apply ideas from linear system theory to neural networks. These problems do not arise in suitably designed nonlinear associative networks (Section III).

The desire to preserve the framework of linear system theory also led Anderson et al. (1977) to employ a homunculus in their model of categorical perception, which cannot adequately be explained by a linear model. To start their discussion of categorical perception, they allowed some of their short-term memory activities to become amplified by positive linear feedback. Left unchecked in a linear model, the positive feedback would force the activities to become infinite, which is physically impossible. To avoid this property, the authors imposed a rule that stops the growth of each activity when it reaches a predetermined maximal or minimal size and thereafter stores this extremal value in memory. The tendency of all variables to reach a maximal or minimal value is then used to discuss data about categorical perception. No physical process is defined to justify the discontinuous change in the slope of each variable when it reaches an extreme of activity or to explain the subsequent storage of these activities. The model thus invokes a homunculus to explain both categorical perception and short-term memory storage.

If the discontinuous saturation rule is replaced by a continuous saturation rule, and if the dynamics of short-term memory storage are explicitly defined, then positive linear feedback can compress the stored activity pattern, rather than contrast enhance it, as one desires to explain categorical perception (Grossberg, 1973, 1978d). This example illustrates how perilous it is to substitute formal algebraic rules, such as those of linear system theory, for dynamical rules in the explication of a psychological process. Even in cases where the algebraic rule seems to express an intuitive property of the psychological process—such as the tendency to saturate—the algebraic rule may also suggest the use of other rules—such as linear positive feedback—that produce diametrically opposed results when they are used in a dynamical description of the process. No homunculus is needed to explain categorical perception in suitably designed nonlinear neural networks (Sections XVIII and XXII). Indeed, nonlinear network mechanisms are designed to avoid the types of instabilities and interpretive anomalies that a linear feedback system approach often generates in a neural network context.

II.F. Classical Conditioning and Attentional Processing

Much as Anderson et al. (1977) improperly lumped all associative models into a Hebbian category, so Sutton and Barto (1981) have incorrectly claimed that associative models other than their own use Hebbian synapses. They go on to reject all Hebbian models in favor of their own non-Hebbian associative model. Given the apparent importance of the Hebbian distinction, it is necessary to define a Hebbian synapse and to analyze why it is being embraced or rejected.

Sutton and Barto (1981, p. 135) follow Hebb to define a Hebbian synapse as follows: "When a cell A repeatedly and persistently takes part in firing another cell B, then A's efficiency in firing B is increased." However, in my associative theory, which Sutton and Barto classify as a Hebbian theory, repeated and persistent associative pairing between A and B can yield conditioned decreases, as well as increases, in synaptic strength (Grossberg, 1969b, 1970b, 1972c). This is not a minor property, since it is needed to assert that the unit of long-term memory is a spatial pattern of synaptic strengths (Section IV). Hebb's law by contrast, is consistent with the assumption that the unit of long term memory is a single synaptic strength. This property does not satisfy the definition of a Hebbian synapse; hence, my associative laws are not Hebbian, contrary to Sutton and Barto's claim. Moreover, the associative component of these laws is only one of several interesting factors that control their mathematical and behavioral properties. None of these factors was considered by Hebb.

Notwithstanding these important details, we still need to ask why Sutton and Barto attack "Hebbian" models. The reason is that Hebbian theories are purported to be unable (1) to recall a conditioned response with a shorter time lag after the presentation of a conditioned stimulus (CS) than was required for efficient learning to occur between the CS and the unconditioned stimulus (UCS), or (2) to explain the inverted U in learning efficacy that occurs as a function of the time lag between a CS and UCS on learning trials. Indeed, Sutton and Barto (1981, p. 142) confidently assert: "Not one of the adaptive element models currently in the literature is capable of producing behavior whose temporal structure is in agreement with that observed in animal learning as described above." Unfortunately, this assertion is false. In fact, Sutton and Barto refer to the article by Grossberg (1974) which reviews a conditioning theory that can explain these phenomena (Grossberg, 1971, 1972a, 1972b, 1975), as well as a variety of other phenomena that Sutton and Barto cannot explain due to their model's formal kinship with the Rescorla-Wagner model (Grossberg, 1982b; Rescorla & Wagner, 1972). Moreover, my explanation does not depend on the non-Hebbian nature of my associative laws but rather on the global anatomy of the networks that I derive to explain conditioning data.

This anatomy includes network regions, called "drive representations," at which the reinforcing properties of external cues join together with internal drive inputs to compute motivational decisions that modulate the attentional processing of external cues. No such concept is postulated in Sutton and Barto's (1981) model. Thus the fact that a pair of simultaneous CSs can be processed, yet a CS that is simultaneous wiih a UCS is not processed, does not depend on the elaboration of the UCS's motivational and attentional properties in the Sutton and Barto model, despite the fact that the UCS might have been a CS just hours before. Sutton and Barto's model of classical conditioning excludes motivational and attentional factors, instead seeking all explanations of classical conditioning data in the properties of a single synapse. Such an approach cannot explain the large database concerning network interactions between neocortex, hypothalamus, septum, hippocampus, and reticular formation in the control of stimulus–reinforcer properties (Berger & Thompson, 1978; Deadwyler, West, & Robinson, 1981; DeFrance, 1976; Gabriel, Foster, Orona, Saltwick, & Stanton, 1980; Haymaker, Anderson, & Nauta, 1969; MacLean, 1970; O'Keefe & Nadel, 1978; Olds, 1977; Stein, 1958; West, Christian, Robinson, & Deadwyler, 1981) and leads its authors to overlook the fact that such interactions are interpreted and predicted by alternative models (Grossberg, 1975). The present chapter also focuses on behavioral properties that are emergent properties of network interactions, rather than of single cells, and illustrates that single cell and network laws must both be carefully chosen to generate desirable emergent properties.

II.G. Search of Associative Memory

The Anderson et al. (1977) model provides one example of a psychological model whose intuitive basis is not adequately instantiated by its formal operations. Such a disparity between intuition and formalism causes internal weaknesses that limit the explanatory and predictive power of many psychological models. These weaknesses can coexist with a model's ability to achieve good data fits on a limited number of experiments. Unfortunately, good curve fits have tended to inhibit serious analysis of the internal structure of psychological models.

Another example of this type of model is Raaijmakers and Shiffrin's (1981) model of associative memory search. The data fits of this model are remarkably good. One reason for its internal difficulties is viewed by the authors as one of its strengths: "Because our main interest lies in the development of a retrieval theory, very few assumptions will be stated concerning the interimage structure" (Raaijmakers & Shiffrin, 1981, p. 123). To characterize this retrieval theory, the model defines learning rules that are analogous to laws of associative learning. However, in information processing models of this kind, terminology like short-term memory (STM) and long-term memory (LTM) is often used instead of terminology like CS, UCS, and conditioning. These differences of terminology seem to have sustained the separate development of models that describe mechanistically related processes.

Although Raaijmakers and Shiffrin's (1981) model intuitively discusses STM and LTM, no STM variables are formally defined; only LTM strengths are defined. This omission forces compensatory assumptions to be made through the remaining theoretical structure. In particular, the LTM strength $S(W_{iT}, W_{jS})$ between the ith word at test (T) and the jth stored (S) word is made a linear function

$$S(W_{iT}, W_{jS}) = bt_{ij} \tag{1}$$

of the time t_{ij} during which both words are in the STM buffer. Thus there is no forgetting, the LTM strength grows linearly to infinity on successive trials, and although both words are supposedly in the buffer when LTM strength is growing, strength is assumed to grow between W_{iT} and W_{iS} rather than between W_{iS} and W_{jS}. A more subtle difficulty is that time per se should not explicitly determine a dynamical process, as it does in Equation 1, unless it parameterizes an external input. All of these problems arise because the theory does not define STM activities which can mediate the formation of long-term memories.

Instead of using STM activities as the variables that control performance, the theory defines sampling and recovery probabilities directly in terms of LTM traces. The sampling probabilities are built up out of products of LTM traces, as in the formula

$$P_S(W_{iS} \mid C_T, W_{kT}) = \frac{S(C_T, W_{iS})S(W_{kT}, W_{iS})}{\Sigma_{j=1}^{n} S(C_T, W_{jS})S(W_{kT}, W_{jS})} \tag{2}$$

for the probability of sampling the ith word W_{iS} given a probe consisting of a context cue C_T and the kth word W_{kT} at test (T). This formula formally compensates for the problem of steadily increasing strengths by balancing numerator strengths against denominator strengths. It also formally achieves selectivity in sampling by multiplying strengths together. The theory does not, however, explain how or why these operations might occur in vivo.

The context cue C_T is of particular importance because the relative strength of context-to-word associations is used to explain the theory's proudest achievement—the part-list cuing effect. However, the context cue is just an extra parameter in the theory because no explanation is given of how a context representation arises or is modified due to experimental manipulations. In other words, because recall theory says nothing about chunking or recognition, the context cue plays a role akin to that played by the "fixed stars" in classical explanations of centrifugal force.

In addition to the continuous rule for strength increase (Equation 1), the theory defines a discrete rule for strength increase

$$S'(W_{iT}, W_{jS}) = S(W_{iT}, W_{jS}) + g$$

which also leads to unbounded strengths as trials proceed. The incrementing rule (Equation 3) is applied only after a successful recall. Although this rule helps to fit some data, it is not yet explained why two such different strengthening rules should coexist.

The authors represent the limited capacity of STM by appending a normalization constraint onto their sampling probability rule. They generalize Equation 2 with the sampling rule

$$P_S(I_i \mid Q_1, Q_2, \ldots, Q_n) = \frac{\Pi_{j=1}^{n} S(Q_j, I_i)^{w_j}}{\Sigma_k \Pi_{j=1}^{n} S(Q_j, I_k)^{w_j}}$$

where the weights w_j satisfy

$$\sum_{j=1}^{m} w_j \leq W.$$

Equation 4 defines the probability of sampling the ith image I_i, given the set of probe cues $Q_1, Q_2, \ldots, Q_n$. Why these normalization weights, which intend to represent the limited capacity of STM, should appear in a sampling rule defined by LTM traces is unexplained in the theory.

The properties that the formalism of Raaijmakers and Shiffrin (1981) attempts to capture have also arisen within my own work on human memory (Grossberg, 1978a, 1978b). Because this theory describes the self-organization of both recognition and recall using real-time operations on STM and LTM traces, it exhibits these properties in a different light. Its analog of the product rule (Equation 2) is due to properties of temporal order information in STM derived from a principle that guarantees the stable transfer of temporal order information from STM to LTM (Section XXXIV). Its analog of the continuous strengthening rule (Equation 1) is found in the chunking process whereby recognition chunks are formed (Section XXI). Its analog of the discrete strengthening rule (3) is due to the process whereby associations from recognition chunks to recall commands are learned (Section VI). Its analog of the normalization rule (Equation 5) is a normalization property of competitive STM networks that are capable of retuning their sensitivity in response to variable operating loads (Section XVII). Not surprisingly, the part-list cuing effect poses no problem for this theory, which also suggests how contextual representations are learned. In light of these remarks, I suggest that Raaijmakers and Shiffrin (1981) have not realized how much the data they wish to explain depends on the "interimage structure" that their theory does not consider.

A few principles and mechanisms based on ideas about self-organization have, in fact, been the vantage point for recognizing and avoiding internal difficulties within psychological models of cognition, perception, conditioning, attention, and information processing (Grossberg, 1978a, 1978e, 1980b, 1980d, 1981b, 1982b, 1982d, 1983, 1984a, 1984b). Some of these principles and mechanisms of self-organization are defined below and used to discuss issues and data concerning the functional units of speech, language, and motor control. This foundation was originally built up for this purpose in Grossberg (1978e). That article, as well as others that derive the concepts on which it is based, are reprinted in Grossberg (1982d).

III. ASSOCIATIVE LEARNING BY NEURAL NETWORKS: INTERACTIONS BETWEEN STM AND LTM

The foundation of the theory rests on laws for associative learning in a neural network, which I call the "embedding field" equations (Grossberg, 1964). These laws are derived from psychological principles and have been physiologically interpreted in many places (e.g., Grossberg, 1964, 1967, 1968b, 1969b, 1970b, 1972c, 1974). They are reviewed herein insofar as their properties shed light on the problem of serial order.

The associative equations describe interactions among unitized nodes v_i that are connected by directed pathways, or *axons* e_{ij}. These interactions are defined in terms of STM traces $x_i(t)$ computed at the nodes v_i and LTM traces z_{ij} computed at the endpoints, or *synaptic knobs*, S_{ij} of the

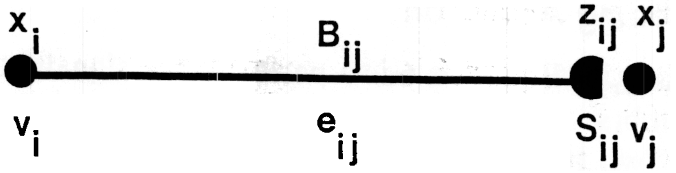

Figure 6.1. An STM trace x_i fluctuates at each node v_i, and an LTM trace z_{ij} fluctuates at the end (synaptic knob) S_{ij} of each conditionable pathway e_{ij}. The performance signal B_{ij} is generated in e_{ij} by x_i and travels at a finite velocity until it reaches S_{ij}. The LTM trace z_{ij} computes a time average of the contiguous STM trace x_j multiplied by a sampling signal E_{ij} that is derived from B_{ij}. The performance signal B_{ij} is gated by z_{ij} before the gated signal $B_{ij}z_{ij}$ perturbs x_j.

directed pathways e_{ij} (Figure 6.1). The simplest realization of these interactions among n nodes $v_1, v_2, \ldots, v_n$ is given by the system of differential equations

$$\frac{d}{dt} x_i = -A_i x_i + \sum_{k=1}^{n} B_{ki} z_{ki} - \sum_{k=1}^{n} C_{ki} + I_i(t)$$

and

$$\frac{d}{dt} z_{ij} = -D_{ij} z_{ij} + E_{ij}[x_j]^+,$$

where $i, j = 1, 2, \ldots, n$; d/dt denotes the rate of change of the contiguous variable, x_i or z_{ij}, as the case might be; and the notation $[\xi]^+ = \max(\xi, 0)$ defines a threshold. The terms in Equations 6 and 7 have the following interpretations.

III.A. STM Decay

Function A_i in Equation 6 is the decay rate of the STM trace x_i. This rate can, in principle, depend on all the unknowns of the system, as in the competitive interaction

$$A_i = A - (B - x_i)g(x_i) + \sum_{k=1}^{n} c_{ik} h(x_k),$$

which I describe more fully in Section XVIII). Equation 8 illustrates that STM decay need not be a passive process. Active processes of competitive signaling, as in this equation or other feedback interactions, can be absorbed into the seemingly innocuous term $A_i x_i$ in Equation 6.

III.B. Spreading Activation

Function B_{ki} in Equation 6 is a performance signal from node v_k to the synaptic knob(s) S_{ki} of pathway e_{ki}. Activation "spreads" along e_{ki} via the signal B_{ki}. Two typical choices of B_{ki} are

$$B_{ki}(t) = b_{ki}[x_k(t - \tau_{ki}) - \Gamma_{ki}]^+ \tag{9}$$

or

$$B_{ki}(t) = f(x_k(t - \tau_{ki}))b_{ki}, \tag{10}$$

where $f(\xi)$ is a sigmoid, or S-shaped, function of ξ with $f(0) = 0$. In Equation 9, a signal leaves v_k only if x_k exceeds the signal threshold Γ_{ki} (Figure 6.2a). The signal moves along e_{ki} at a finite velocity ("activation spreads") and reaches S_{ki} after τ_{ki} time units. Typically, τ_{ki} is a short time compared to the time it takes v_k to exceed threshold Γ_{ki} in response to signals. Parameter b_{ki} measures the strength of the pathway e_{ki} from v_k to v_i. If $b_{ki} = 0$, no pathway exists.

In Equation 10, the signal threshold Γ_{ki} is replaced by attenuation of the signal at small x_k values and by saturation of the signal at large x_k values (Figure 6.2b). The S-shaped signal function is the simplest physical signal function that can prevent noise amplification from occurring due to reverberatory signaling in a feedback network (Section XVIII).

III.C. Probed Read-Out of LTM: Gating of Performance Signals

Term $B_{ki}z_{ki}$ in Equation 6 says that the signal B_{ki} from v_k to S_{ki} interacts with the LTM trace z_{ki} at S_{ki}. This interaction can be intuitively described in several ways. For one, B_{ki} is a probe signal, activated by STM at v_k, that reads out the LTM trace z_{ki} into the STM trace x_i of v_i. For another, z_{ki} "gates" signal B_{ki} before it reaches v_i from v_k so that the signal strength that perturbs x_i at v_i is $B_{ki}z_{ki}$ rather than B_{ki}. Thus even if an input to v_k excited equal signals B_{ki} in all the pathways e_{ki}, only those v_i abutted by large LTM traces z_{ki} will be appreciably activated by v_k. Activation does not merely "spread" from v_k to other nodes; it can be transformed into propagated signals (x_k into B_{ki}) and gated by LTM traces (B_{ki} into $B_{ki}z_{ki}$) before it reaches these nodes.

Adaptive Filtering

The gated signals from all the nodes v_k combine additively at v_i to form the total signal $T_i = \Sigma_{k=1}^{n} B_{ki}z_{ki}$ of Equation 6. Speaking mathematically, T_i is the "dot product," or inner product, of the vectors $B_i = (B_{1i}, B_{2i},$

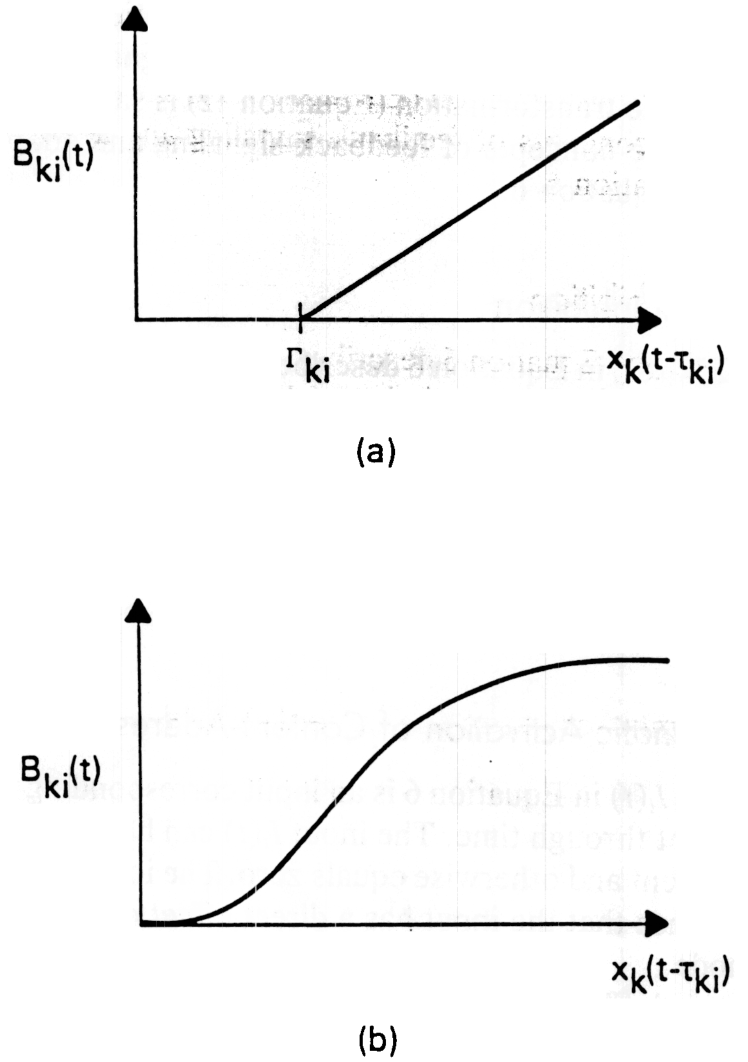

Figure 6.2 (a) A threshold signal: $B_{ij}(t)$ is positive only if $x_i(t - \tau_{ij})$ exceeds the signal threshold Γ_{ij}. $B_{ij}(t)$ is a linear function of $x_j(t - \tau_{ij})$ above this threshold. (b) A sigmoid signal: $B_{ij}(t)$ is attenuated at small values of $x_i(t - \tau_{ij})$, much as in the threshold case, and levels off at large values of $x_i(t - \tau_{ij})$ after all signaling sites are turned on.

. . . , B_{ni}) and $z_i = (z_{1i}, z_{2i}, \ldots, z_{ni})$ of probe signals and LTM traces, respectively. Such a dot product is often written as

$$T_i = B_i \cdot z_i. \tag{11}$$

The transformation of the vector $x^* = (x_1, x_2, \ldots, x_n)$ of all STM traces into the vector $T^* = (T_1, T_2, \ldots, T_n)$ of all dot products, specifically

$$x^* \rightarrow T^*,$$

completely describes how STM traces generate feedback signals within the network.

A transformation by dot products as in Equation 12 is said to define a *filter*. Because the LTM traces z_i that gate the signals B_i can be changed by experience, the transformation (Equation 12) is said to define an *adaptive filter*. Thus the concepts of feedback signaling and adaptive filtering are identical in Equation 6.

III.E. Lateral Inhibition

Term $\Sigma_{k=1}^{n} C_{ki}$ in Equation 6 describes the total inhibitory signal from all nodes v_k to v_i. An illustrative choice of the inhibitory signal from v_k to v_i is

$$C_{ki}(t) = g(x_k(t - \sigma_{ki}))c_{ki}, \tag{13}$$

where $g(\xi)$ is a sigmoid signal function, σ_{ki} is the time lag for a signal to be transmitted ("spread") between v_k and v_i, and c_{ki} describes the strength of the inhibitory path from v_k to v_i.

Automatic Activation of Content-Addressable Nodes

Function $I_i(t)$ in Equation 6 is an input corresponding to presentation of the ith event through time. The input $I_i(t)$ can be large during and shortly after the event and otherwise equals zero. The input automatically excites v_i in the sense that the input has a direct effect on the STM activity of its target node.

In all, each STM trace can decay, can be activated by external stimuli, and can interact with other nodes via sums of gated excitatory signals and inhibitory signals. These equations can be generalized in several ways (Grossberg, 1974, 1982d). For example, LTM traces for inhibitory pathways can also be defined (Grossberg, 1969b) and in a way that avoids the difficulties of Estes's (1972) theory in Section I. The appendix describes a more general version of the equations that includes stable, conditionable inhibitory pathways.

III.G. LTM Decay

Function D_{ij} in Equation 7 is the decay rate of the LTM trace z_{ij}. The LTM decay rate, like the STM decay rate, can depend on the state of the system as a whole. For example, in principle it can be changed by attentional signals, probe signals, slow threshold fluctuations, and the like without destroying the invariants of associative learning that I need to carry out my argument (Grossberg, 1972c, 1974, 1982d).

III.H. Read-In of STM into LTM: Stimulus Sampling

Function E_{ij} in Equation 7 describes a learning signal from v_i to S_{ij} that drives the LTM changes in z_{ij} at S_{ij}. In other words, v_i "samples" v_j by turning on E_{ij}. Otherwise expressed, the STM trace x_j is read into the LTM trace z_{ij} by turning on the sampling signal E_{ij}. In the simplest cases, E_{ij} is proportional to B_{ij}. By setting both D_{ij} and E_{ij} equal to zero in Equation 7, a pathway e_{ij} can be converted from a conditionable pathway to a prewired pathway that is incapable of learning.

An important technical issue concerns the most general relationship that can exist between B_{ij} and E_{ij}. It has been proven that, in a precise mathematical sense, unbiased learning occurs if "B_{ij} is large only if E_{ij} is large" (Grossberg, 1972c, 1982d). This condition, called a "local flow" condition, is interpreted physically as follows. After the sampling signal E_{ij} reaches S_{ij}, it influences learning by z_{ij} within S_{ij}. The sampling signal E_{ij} is also averaged, delayed, or otherwise transformed within S_{ij} to give rise to the performance signal B_{ij}. This signal acts at a "later stage" within S_{ij} than E_{ij} because B_{ij} energizes the net effect $B_{ij}z_{ij}$ of v_i on v_j. The mathematical local flow condition shows that this physical interpretation of the relationship between E_{ij} and B_{ij} is sufficient to guarantee unbiased learning.

III.I. Mutual Interaction of STM and LTM

By joining together terms $D_{ij}z_{ij}$ and $E_{ij}x_j$, it follows from Equation 7 that the LTM trace z_{ij} is a time average of the product of learning signals E_{ij} from v_i to S_{ij}, with STM traces at v_j. When z_{ij} changes in size, it alters the gated signals from v_i to v_j via term $B_{ij}z_{ij}$, and thus the value of the STM trace x_j. In this way the STM and LTM traces mutually influence each other, albeit on different spatial and temporal scales.

IV. LTM UNIT IS A SPATIAL PATTERN: SAMPLING AND FACTORIZATION

To understand the functional units of goal-oriented behavior, it is necessary to characterize the functional unit of long-term memory in an associative network. This problem was approached by first analyzing what the minimal anatomy capable of associative learning can actually learn (Grossberg, 1967, 1968a, 1969g, 1970b) and then proving that the same functional unit of memory is computed in much more general anatomies (Grossberg, 1969b, 1972c, 1974). Three properties that were discovered by these investigations will be needed here:

1. The functional unit of LTM is a *spatial pattern* of activity.
2. A spatial pattern is encoded in LTM by a process of *stimulus sampling.*
3. The learning process *factorizes* the input properties which energize learning and performance from the spatial patterns to be learned and performed.

Each of these abstract properties is a computational universal that appears under different names in ostensibly unrelated concrete applications. Henceforth in the chapter, an abstract property will be described before it is applied to concrete examples.

V. OUTSTAR LEARNING: FACTORIZING COHERENT PATTERN FROM CHAOTIC ACTIVITY

The minimal anatomy capable of associative learning is depicted in Figure 6.3a. A single node, or population, v_0 is activated by an external event via an input function $I_0(t)$. This event is called the *sampling event.* For example, in studies of classical conditioning, the sampling event is the conditioned stimulus (CS).

If the sampling event causes the signal thresholds of node v_0 to be exceeded by its STM trace x_0, then learning signals E_{0i} propagate along the pathways e_{0i} toward a certain number of nodes v_i, $i = 1, 2, \ldots, n$. The same analysis of learning applies no matter how many nodes v_i exist, provided that at least two nodes exist ($n \geq 2$) to permit some learning to occur. The learning signals E_{0i} are also called *sampling signals* because their size influences the learning rate, with no learning occurring when all signals E_{0i} are equal to zero.

The sampling signals E_{0i} from v_0 do not activate the nodes v_i directly. In contrast, the LTM-gated performance signals $B_{0i}z_{0i}$ directly influence the nodes v_i by activating their STM traces x_i. The nodes v_i can also be activated directly by the events to be learned. These events are represented by the inputs $I_i(t)$ which activate the STM traces x_i of the nodes v_i, $i = 1, 2, \ldots, n$. Because the signals E_{0i} enable the z_{0i} to sample STM traces, the inputs $I_1(t), I_2(t), \ldots, I_n(t)$ are called the *sampled event.* In studies of classical conditioning, the sampled event is the unconditioned stimulus (UCS). The output signals from the nodes v_i that are caused by the UCS control the network's unconditioned response (UCR).

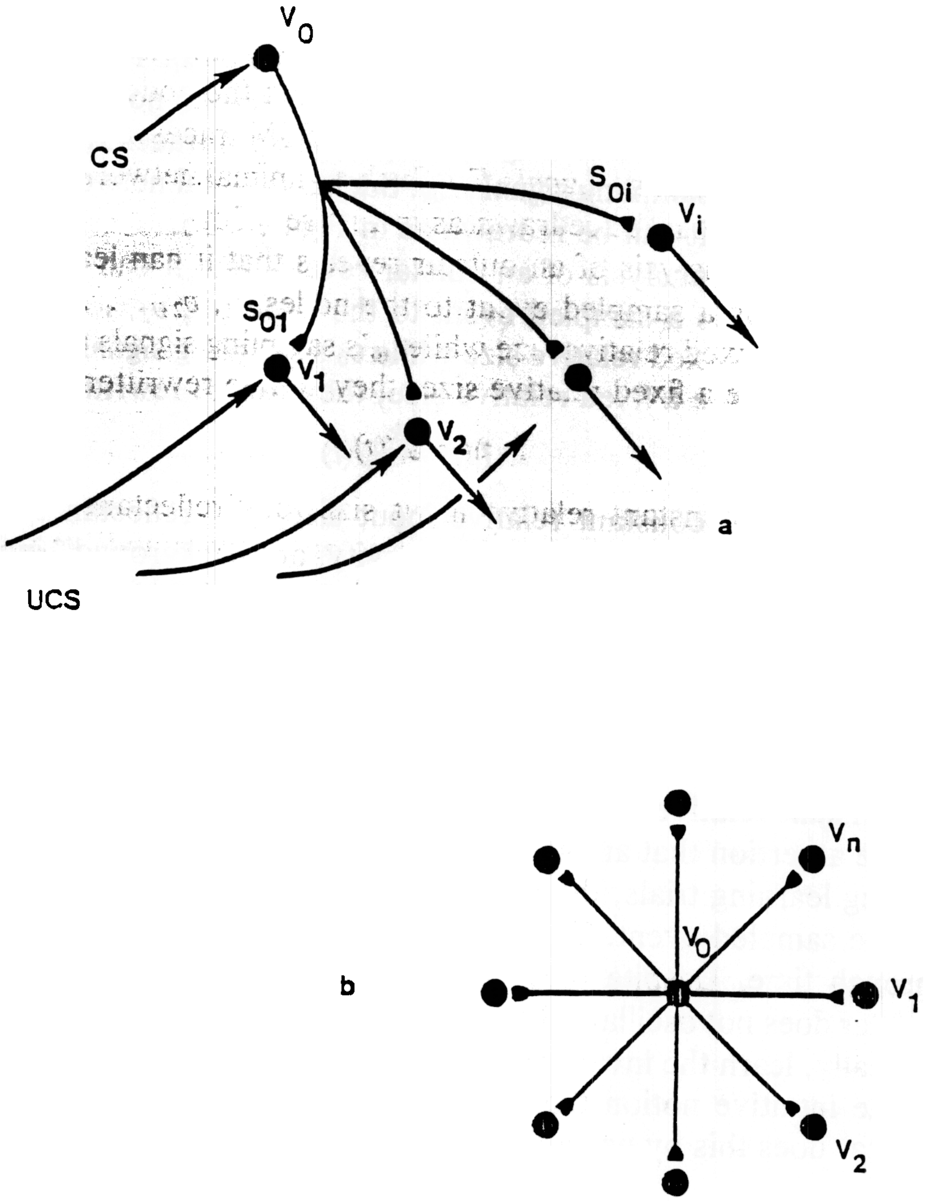

Figure 6.3 The minimal network capable of associative pattern learning: (a) A conditioned stimulus (CS) activates a single node or cell population v_0, which sends sampling signals to a set of nodes $v_1, v_2, \ldots, v_n$. An input pattern representing an unconditioned stimulus (UCS) activates the nodes $v_1, v_2, \ldots, v_n$, which elicit output signals that contribute to the unconditioned response (UCR). The sampling signals from v_0 activate the LTM traces z_{0i} that are computed at the synaptic knobs S_{0i}, $i = 1, 2, \ldots, n$. The activated LTM traces can learn the activity pattern across $v_1, v_2, \ldots, v_n$ that represents the UCS. (b) When the sampling network in (a) is drawn to emphasize its symmetry, the result is an *outstar* wherein v_0 is the sampling source and the set $\{v_1, v_2, \ldots, v_n\}$ is the sampled border.

The sampling signals E_{0i} directly activate the performance signals B_{0i} and the LTM traces z_{0i} rather than the STM traces x_i. These LTM traces are computed at the synaptic knobs S_{0i} that abut the nodes v_i. This location permits the LTM traces z_{0i} to sample the STM traces x_i when they are activated by the sampling signals E_{0i}. Such a minimal network is called an *outstar* because it can be redrawn as in Figure 6.3b.

Mathematical analysis of an outstar reveals that it can learn a *spatial pattern,* which is a sampled event to the nodes $v_1, v_2, \ldots, v_n$ whose inputs I_i have a fixed relative size while v_0's sampling signals are active. If the inputs I_i have a fixed relative size, they can be rewritten in the form

$$I_i(t) = \theta_i I(t) \tag{14}$$

where θ_i is the constant relative input size, or "reflectance," and the function $I(t)$ is the fluctuating total activity, or "background" input, of the sampled event. The convention that $\Sigma_{i=1}^{n} \theta_i = 1$ guarantees that $I(t)$ represents the total sampled input to the outstar—specifically, $I(t) = \Sigma_{i=1}^{n} I_i(t)$. The pattern weights of the sampled event is the vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_{\mathrm{n}})$$

of constant relative input sizes. The outstar learns this vector.

The assertion that an outstar can learn a vector θ means the following. During learning trials, the sampling event is followed a number of times by the sampled event. Thus the inputs $I_0(t)$ and $I(t)$ can oscillate wildly through time. Despite these wild oscillations, however, learning in an outstar does not oscillate. Rather, the outstar can progressively, or monotonically, learn the invariant spatial pattern θ across trials, corresponding to the intuitive notion that "practice makes perfect" (Figure 6.4). The outstar does this by using the fluctuating inputs $I_0(t)$ and $I(t)$ as energy to

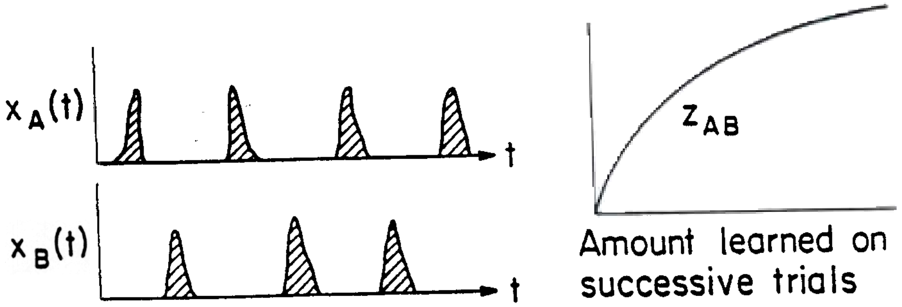

Figure 6.4 Oscillatory inputs due to repetitive A-then-B presentations are translated into a monotonic learned reaction of the corresponding stimulus sampling probabilities. In the text, fluctuations in the sampling input $I_0(t)$ and total sampled input $I(t)$, as well as the monotonic reactions of the relative LTM traces $Z_{0i}(t)$, generalize the A-then-B interpretation.

drive its encoding of the pattern θ. The fluctuating inputs $I_0(t)$ and $I(t)$ determine the rate of learning but not the pattern θ that is learned. This is the property of *factorization*: fluctuating input energy determines learning rate, while the invariant input pattern determines what is learned. The factorization property shows that the outstar can detect and encode temporally coherent relationships among the inputs that represent the sampled event.

In mathematical terms, factorization implies that the relative LTM traces

$$Z_i = z_{0i}\left(\sum_{k=1}^{n} z_{0k}\right)^{-1}$$

are drawn monotonically toward the target ratios θ_i. Stimulus sampling means that the LTM ratios Z_i change only when the sampling signals from v_0 to the synaptic knobs S_{0i} are positive. Because the LTM ratios form a probability distribution (each $Z_i \geq 0$ and $\Sigma_{i=1}^{n} Z_i = 1$) and change only when sampling signals are emitted, I call them the *stimulus sampling probabilities* of an outstar. The behavior of these quantities explicates the probabilistic intuitions underlying stimulus sampling theory (Neimark & Estes, 1967) in terms of the deterministic learning dynamics of a neural network. In particular, the factorization property dynamically explains various properties that are assumed in a stimulus sampling model—for example, why learning curves should be monotonic in response to wildly oscillating inputs (Figure 6.4).

The property of factorization also has an important meaning during performance trials. Both sampling signals and performance signals are released during performance trials (Grossberg, 1972c). The property of factorization means that the performance signal may be chosen to be any nonnegative and continuous function of time without destroying the outstar's memory of the spatial pattern that was encoded in LTM on learning trials. The main constraint is that the pattern weights θ_i be read out synchronously from all the nodes v_i.

What happens if the sampled event to an outstar is not a spatial pattern, as in the case when a series of sampled events occur, rather than a single event? Such an event series can be represented by a vector input

$$J(t) = (I_1(t), I_2(t), \ldots, I_n(t)), \tag{17}$$

$t \geq 0$, where each input $I_i(t)$ is a nonnegative and continuous function of time. Because each input $I_i(t)$ is continuous, the relative pattern weights

$$\theta_i(t) = I_i(t)\left[\sum_{k=1}^{n} I_k(t)\right]$$

are also continuous functions of time, as is the vector function

$$\theta(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_n(t)) \tag{19}$$

of pattern weights.

Mathematical analysis of the outstar reveals that its LTM traces learn a spatial pattern even if the weights $\theta(t)$ vary through time. The spatial pattern that is encoded in LTM is a weighted average of all the spatial patterns $\theta(t)$ that are registered at the nodes v_i while sampling signals from v_0 are active.

This result raises the question, How can each of the patterns $\theta(t)$ be encoded in LTM rather than an average of them all? The properties of outstar learning (Section IV) readily suggest an answer to this question. This answer propelled the theory on one of its roads toward a heightened understanding of the serial order problem. Before following this road, some applications of outstar learning are now summarized.

VI. SENSORY EXPECTANCIES, MOTOR SYNERGIES, AND TEMPORAL ORDER INFORMATION

The fact that associative networks encode spatial patterns in LTM suggests that the brain's sensory, motor, and cognitive computations are all pattern transformations. This expectation arises from the fact that computations which cannot in principle be encoded in LTM can have no adaptive value and thus would presumably atrophy during evolution. Examples of spatial patterns as functional units of sensory processing include the reflectance patterns of visual processing (Cornsweet, 1970), the sound spectrograms of speech processing (Cole, Rudnicky, Zue, & Reddy, 1980; Klatt, 1980), the smell-induced patterns of olfactory bulb processing (Freeman, 1975), and the taste-induced patterns of thalamic processing (Erickson, 1963). More central types of pattern processing are also needed to understand the self-organization of serial order.

VI.A. Sensory Expectancies

Suppose that the cells $v_1, v_2, \ldots, v_n$ are sensory feature detectors in a network's sensory cortex. A spatial pattern across these feature detectors may encode a visual or auditory event. The relative activation of each v_i then determines the relative importance of each feature in the global STM representation of the event across the cortex. Such a spatial pattern code can effectively represent an event even if the individual feature detectors v_i are broadly tuned. Using outstar dynamics, even a single command node v_0 can learn and perform an arbitrary sensory representation of this sort. The pattern read out by v_0 is often interpreted as the representation that v_0 "expects" to find across the field $v_1, v_2, \ldots, v_n$ due to prior experience. In this context, outstar pattern learning illustrates top-down expectancy learning (Section XXV). The expectancy controlled by a given node v_0 is a time average of all the spatial patterns that it ever sampled. Thus it need not equal any one of these patterns.

VI.B. Motor Synergies

Suppose that the cells $v_1, v_2, \ldots, v_n$ are motor control cells such that each v_i can excite a particular group of muscles. A larger signal from each v_i then causes a faster contraction of its target muscles. Spatial pattern learning in this context means that an outstar command node v_0 can learn and perform fixed relative rates of contraction across all the motor control cells $v_1, v_2, \ldots, v_n$. Such a spatial pattern can control a motor synergy, such as playing a chord on the piano with prescribed fingers; making a synchronous motion of the wrist, arm, and shoulder; or activating a prescribed target configuration of lips and tongue while uttering a speech sound (Section XXXII).

Because outstar memory is not disturbed when the performance signal from v_0 is increased or decreased, such a motor synergy, once learned, can be performed at a variety of synchronous rates without requiring the motor pattern to be relearned at each new rate. (Kelso, Southard, & Goodman, 1979; Soechting & Laquaniti, 1981). In other words, the factorization of pattern and energy provides a basis for independently processing the command needed to reach a terminal motor target and the velocity with which the target will be approached.

This property may be better understood through the following example. When I look at a nearby object, I can choose to touch it with my left hand, my right hand, my nose, and so on. Several *terminal motor maps* are simultaneously available to move their corresponding motor organs towards the object. "Willing" one of these acts releases the corresponding terminal motor map but not the others. The chosen motor organ can, moreover, be moved toward the invariant goals at a wide range of velocities. The distinction between the invariant terminal motor map and the flexibility programmable performance signal illustrates how factorization prominently enters problems of learned motor control.

VI.C. Temporal Order Information over Item Representations

Suppose that a sequence of item representations is activated in a prescribed order during perception of a list. At any given moment, a spatial pattern of STM activity exists across the excited populations. Were thc same items excited in a different order by a different list, a different spatial pattern of STM activity would be elicited. Thus the spatial pattern reflects temporal order information as well as item information. An outstar sampling source can encode this spatial pattern as easily as any other spatial pattern. Thus, although an outstar can encode only a spatial pattern, this pattern can represent temporal properties of external events. Such a spatial encoding of temporal order information is perhaps the example par excellence of a network's parallel processing capabilities. How a network can encode temporal order information in STM without falling into the difficulties mentioned in Section II will be described in Section XXXIV.

VII. RITUALISTIC LEARNING OF SERIAL BEHAVIOR: AVALANCHES

The following sections approach the problem of serial order in stages. These stages mark different levels of sophistication in a network's ability to react adaptively to environmental feedback. The stages represent a form of conceptual evolution in the theory reflecting the different levels of behavioral evolution that exist across phylogeny.

The first stage shows how outstar learning capabilities can be used to design a minimal network capable of associatively learning and/or performing an arbitrary sequence of events, such as a piano sonata or a dance. This construction is called an *avalanche* (Grossberg, 1969g, 1970a, 1970b) because its sampling signal traverses a long axon that activates regularly spaced cells (Figure 6.5) in a manner reminiscent of how avalanche conduction along the parallel fibers in the cerebellum activates regularly spaced Purkinje cells (Eccles, Ito, & Szentagothai, 1967; Grossberg, 1969d). The simplest avalanche requires only one node to encode the memory of the entire sequence of events. Thus the construction shows that complex performance per se is easily achieved by a small and simple neural network.

The simplest avalanche also exhibits several disadvantages stemming from the fact that its performance is ritualistic in several senses. Each of these disadvantages has a remedy that propels the theory forward. Performance is temporally ritualistic because once performance has been initiated, it cannot be rhythmically modified by the performer or by conflicting environmental demands. Performance is spatially ritualistic in the sense that the motor patterns to be performed do not have learned sensory referents.

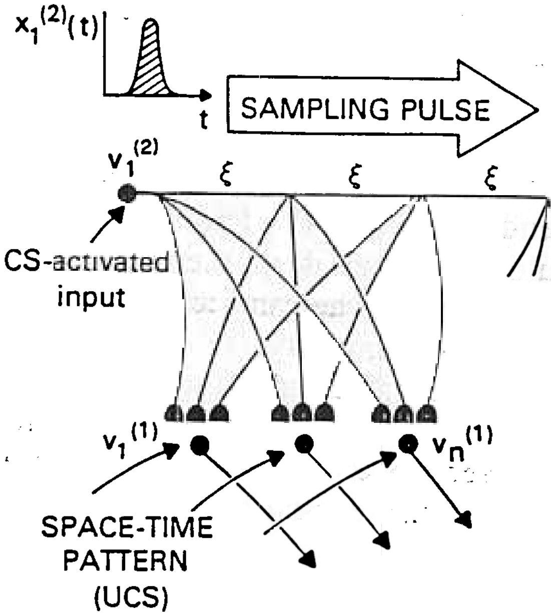

Figure 6.5 An avalanche is the minimal network that can associatively learn and ritualistically perform any space–time pattern. The sampling node $v_1^{(2)}$ emits a brief sampling pulse that serially excites the outstar sampling bouquets that converge on the sampled field $\mathcal{F}^{(1)} = \{v_1^{(1)}, v_2^{(1)}, \ldots, v_n^{(1)}\}$. On performance trials, a sampling pulse resynthesizes the space–time pattern as a series of smoothly interpolated spatial patterns.

The first modification of the avalanche enables performance speed to be continuously modulated, or even terminated, in response to performer or environmental signals. This flexibility can be achieved on performance trials without any further learning of the ordered patterns themselves. The construction thus provides a starting point for analyzing how order information and rhythm can be decoupled in more complex learning situations. The construction is not of merely formal interest, however, since it shares many properties with the command cell anatomies of invertebrates (Dethier, 1968; Hoyle, 1977; Kennedy, 1968; Stein, 1971; Willows, 1968).

With the modified avalanche construction before us, some design issues become evident concerning how to overcome the network's spatially ritualistic properties. The pursuit of these issues leads to a study of serial learning and chunking that in turn provides concepts for building a theory of recognition and recall. The needed serial learning and chunking properties are also properties of the embedding field equations, albeit in a different processing context than that of outstar learning.

Because the avalanche constructions require a hierarchy of network stages, superscripts are used on the following variables. Suppose that the act to be learned is controlled by a set of nodes $v_1^{(1)}, v_2^{(1)}, \ldots, v_n^{(1)}$, henceforth called the field of cells $\mathcal{F}^{(1)}$. This field replaces the nodes $v_1, v_2, \ldots, v_n$ of an outstar. Let each node receive a nonnegative and continuous input $I_i(t)$, $t \geq 0$, $i = 1, 2, \ldots, n$. The set of inputs $I_i(t)$ collectively form a vector input

$$J(t) = (I_1(t), I_2(t), \ldots, I_n(t)),$$

$t \geq 0$, that characterizes the commands controlling the sequence of events. At the end of Section IV, I raised the question of how such a vector input could be learned despite the outstar's ability to learn only one spatial pattern. An avalanche can accomplish this task using a single encoding cell in the following way.

Speaking intuitively, $J(t)$ describes a moving picture playing through time on the "screen" of nodes $\mathcal{F}^{(1)}$. An avalanche can learn and perform such a "movie" as a sequence of still pictures that are smoothly interpolated through time. Because each input $I_i(t)$ is continuous, the pattern weights

$$\theta_i(t) = I_i(t)\left[\sum_{k=1}^{n} I_k(t)\right]^{-1}$$

are also continuous and can therefore be arbitrarily closely approximated by a sequence of values

$$\theta_i(0),\ \theta_i(\xi),\ \theta_i(2\xi),\ \theta_i(3\xi),$$

sampled every ξ time units, if ξ is chosen so small that $\theta_i(t)$ does not change too much in a time interval of length ξ. For every fixed k, the vector of weights

$$\theta^{(k)} = (\theta_1(k\xi), \theta_2(k\xi), \ldots, \theta_n(k\xi)) \tag{21}$$

sampled across all the cells in $\mathcal{F}^{(1)}$ at a time $t = k\xi$ is a spatial pattern. To learn and perform the movie $J(t)$, $t \geq 0$, it suffices to learn and perform the sequence $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ of spatial patterns in the correct order. This can be done if a sequence of outstars $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \ldots$ is arranged so that the kth outstar $\mathcal{O}_k$ samples only the kth spatial pattern $\theta^{(k)}$ on successive learning trials and is then briefly activated in the order $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \ldots$ on performance trials. Using the outstar properties of stimulus sampling and learning in spatial pattern units, an avalanche-type anatomy, such as that in Figure 6.5, instantiates these properties using a single sampling node and a minimum number of pathways.

In Figure 6.5, a brief sampling signal travels along the long pathway (axon) leading from node $v_1^{(2)}$. This node replaces the outstar sampling source v_0 of the previous discussion. This sampling signal moves from left to right, traveling down the serially arranged bouquets of pathways that converge on $\mathcal{F}^{(1)}$. Each bouquet is an outstar and can therefore learn a spatial pattern. This pattern is a weighted average of all the spatial patterns that are active in STM across $\mathcal{F}^{(1)}$ while the outstar samples $\mathcal{F}^{(1)}$. Because each outstar $\mathcal{O}_k$ samples $\mathcal{F}^{(1)}$ briefly, its LTM traces encode essentially only the pattern $\theta^{(k)}$. By the property of stimulus sampling, none of the other patterns $\theta^{(j)}$, $j \neq k$, playing across $\mathcal{F}^{(1)}$ through time will influence the LTM traces of $\mathcal{O}_k$. On performance trials, a performance signal runs along the axon, serially reading out the spatial patterns encoded by the bouquets in the correct order. The STM traces across $\mathcal{F}^{(1)}$ smoothly interpolate these spatial patterns through time to generate a continuously varying output from $\mathcal{F}^{(1)}$.

VIII. DECOUPLING ORDER AND RHYTHM: NONSPECIFIC AROUSAL AS A VELOCITY COMMAND

Performance by an avalanche is temporally ritualistic, because once a performance signal is emitted by node $v_1^{(2)}$, there is no way to temporally modulate or stop its inexorable transit down the sampling pathway. In order to modify performance at any time during the activation of an avalanche, there must exist a locus at each outstar's sampling source where auxiliary inputs can modify the performance signal. These loci are denoted by nodes $v_1^{(2)}$, $v_2^{(2)}$, $v_3^{(2)}$, . . . such that node $v_k^{(2)}$ is the sampling source of outstar $\mathcal{O}_k$, as in Figure 6.6a. These nodes form a field of nodes $\mathcal{F}^{(2)}$ in their own right.

Merely defining the nodes $\mathcal{F}^{(2)}$ does not make avalanche performance less ritualistic as long as a signal from the ith node $v_i^{(2)}$ can trigger a signal from the $(i + 1)$st node $v_{i+1}^{(2)}$. An auxiliary input source (or sources) needs to be defined whose activity is required to maintain avalanche performance. The minimal solution is to define a single node $v_1^{(3)}$ that can activate all the populations in $\mathcal{F}^{(2)}$ (approximately) simultaneously (Figure 6.6b) and to require that node $v_{i+1}^{(2)}$ can elicit a signal only if it receives simultaneous signals from $v_i^{(2)}$ and $v_1^{(3)}$. Shutting off $v_1^{(3)}$ at any time can then abruptly terminate performance, because even if node $v_{i+1}^{(2)}$ receives a signal from $v_i^{(2)}$ at such a time, $v_{i+1}^{(2)}$ cannot emit a signal without convergent input from $v_1^{(3)}$.

Since the signals from $v_1^{(3)}$ energize the avalanche as a whole, I call them *nonspecific arousal* signals. Node $v_1^{(3)}$ may also be called a *command node* because it subliminally prepares the entire avalanche for activation. This node may just as well be thought of as a context node, or contextual bias, in situations wherein the avalanche is interpreted as a subnetwork within a larger network. For example, the same cue can often trigger different behavior depending on the context or plan according to which it is interpreted (e.g., turning left at the end of a hall to enter the

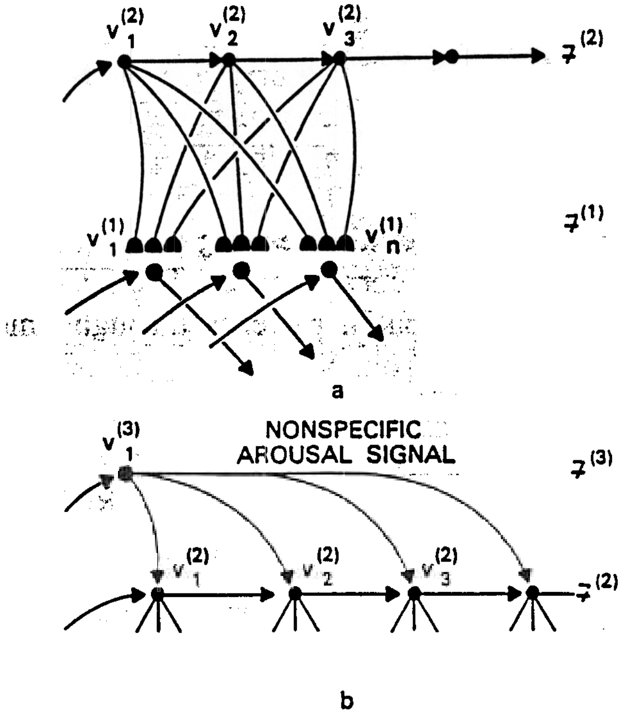

Figure 6.6 A nonspecific arousal signal can act as a command that decouples order information from the velocity or rhythm of a particular performance. (a) Outstar sampling sources $v_j^{(2)}$ serially excite each other to determine the order with which $\mathcal{F}^{(1)}$ is sampled. (b) Simultaneous input from the nonspecific arousal source $v_1^{(3)}$ and $v_i^{(2)}$ is needed to elicit an output signal from $v_{i+1}^{(2)}$. Continuous variations in the size of the $v_1^{(3)}$ signal are translated into continuous variations in the velocity of performance.

bedroom or right to enter the dining room). A nonspecific arousal source can achieve this result by subliminally sensitizing certain subnetworks more than others for supraliminal activation by a given set of cues. Due to the fact that the arousal source often encodes a contextual cue, I henceforth call the network in Figure 6.6 a *context-modulated avalanche*.

IX. REACTION TIME AND PERFORMANCE SPEED-UP

The requirement that the command node $v_{i+1}^{(2)}$ can fire only if it receives simultaneous signals from $v_i^{(2)}$ and $v_1^{(3)}$ can be implemented in several related ways, all of which depend on the existence of a threshold Γ that each STM trace $x_{i+1}^{(2)}$ must exceed before a signal from $v_{i+1}^{(2)}$ can be emitted. Increasing the arousal signal from $v_1^{(3)}$ to $v_{i+1}^{(2)}$ makes it easier for $x_{i+1}^{(2)}$ to exceed this threshold. Moreover, the threshold of $v_{i+1}^{(2)}$ is exceeded faster in response to a large signal from $v_1^{(3)}$ than to a small signal from $v_1^{(3)}$. The reaction time (RT) for activating any node $v_{i+1}^{(2)}$ can thus be decreased by increasing the nonspecific arousal level at the time when this node is excited by $v_i^{(2)}$. Since this is true for every node $v_{i+1}^{(2)}$, a continuous variation of arousal level through time can continuously modulate performance speed. Indeed, such a feedforward modulation of performance velocity can be achieved at a very rapid rate (Lashley, 1951).

Both additive and multiplicative (shunting) rules can, in principle, be used to restrict avalanche performance except when the nonspecific arousal source is active. A shunting rule enjoys an important formal advantage. It works even without an exquisitely precise choice of the relative sizes of all the avalanche's signals and thresholds. A shunting rule has this advantage because a zero arousal level will gate to zero even a large signal from $v_i^{(2)}$ to $v_{i+1}^{(2)}$. Consequently, $v_{i+1}^{(2)}$ will not be activated at all by $v_i^{(2)}$. Its zero STM trace $x_{i+1}^{(2)}$ will therefore remain smaller than any choce of a positive threshold.

The concept of a shunting arousal signal is illustrated by the following equation for the activation of $v_{i+1}^{(2)}$ by $v_i^{(2)}$ and $v_1^{(3)}$:

$$\frac{d}{dt} x_{i+1}^{(2)} = -Ax_{i+1}^{(2)} + B_iS \tag{22}$$

where B_i is the performance signal from $v_i^{(2)}$ and S is the shunting signal from $v_1^{(3)}$. Often B_i has the form $B_i(t) = f(x_i^{(2)}(t - \xi))$ and S has the form $S(t) = g(x_1^{(3)}(t - \tau))$ where both $f(w)$ and $g(w)$ are sigmoid functions of w. For simplicity, let B_i and S equal zero at times $t < 0$ and equal constant positive values U and V, respectively, at times $t \geq 0$. Define the RT of $x_{i+1}^{(2)}(t)$ as the first time $t = T$ when $x_{i+1}^{(2)}(T) = \Gamma$ and $(d/dt)x_{i+1}^{(2)}(T) > 0$. By Equation 22, if $V = 0$, then the RT is infinite since $x_{i+1}^{(2)} \equiv 0$. Only if $UV > A\Gamma$ does a signal ever leave $v_{i+1}^{(2)}$, and it does so with an RT of

$$\frac{1}{A} \ln \left(\frac{UV}{UV - A\Gamma}\right),$$

which is a decreasing function of U and V. Equation 23 shows that determination of performance rate depends on the threshold of each node, the arousal level when the node exceeds threshold, the size of the signal generated by the previous node, and the delays ξ and τ due to propagating signals between nodes.

In more complex examples, the signals to the nodes $v_i^{(2)}$ can also be altered by LTM traces that gate these signals (Section XIII). Equation 23 illustrates how an increase in such an LTM trace due to learning can cause a progressive speed-up of performance (Fitts & Posner, 1967; Welford, 1968), because an increase in an LTM trace that gates either signal B_i or S in the equation is equivalent to an increase in S.

An interaction between learning, forgetting, and reaction time was used to explain such phenomena as performance speed-up due to learning as well as some backward-masking properties in Grossberg (1969c). Later articles on speed-up have used the same ideas to explain the power law relationship that often obtains between the time it takes to perform a task and the number of practice trials (Anderson, 1982; MacKay, 1982). Although both Anderson and MacKay suggest that their explanation of speed-up lends special support to their other concepts, it is historically more correct to say that both explanations lend support to well-known neural modeling ideas.

There exist many variations on the avalanche theme. Indeed, it may be more appropriate to talk about a theory of avalanches than of assorted avalanche examples (Grossberg, 1969g, 1970b, 1974, 1978e). For present purposes, we can summarize some avalanche ideas in a way that generalizes to more complex situations:

1. A source of nonspecific arousal signals can modulate the performance rhythm.
2. Specific activations encode temporal order information in a way that enables a nonspecific arousal signal to generate temporally ordered performance.
3. The sequential events to be learned and performed are decomposed into spatial patterns.

X. HIERARCHICAL CHUNKING AND THE LEARNING OF SERIAL ORDER

Given the concept of a context-modulated avalanche, the following question generates a powerful teleological pressure for extending the theory. Suppose that the links $v_i^{(2)} \rightarrow v_{i+1}^{(2)}$ in the chain of nodes are not prewired into the network. How can a network with the processing capabilities of a context-modulated avalanche be self-organized? For example, if the spatial patterns across the sampled nodes $\mathcal{F}^{(1)}$ control successive chords in a piano sonata, then the individual chords need to be learned as well as the ordering of the chords. This process is typically carried out by reading piano music and transforming these visual cues into motor commands which elicit auditory feedback. Nowhere in a context-modulated avalanche are visual cues encoded before being mapped into motor commands. No sequence of auditory feedback patterns is encoded before being correlated with its generative sequence of motor commands, and no mechanism is included whereby the serial order of the motor commands can be learned.

It is clear from this and many analogous examples that the set of sampling nodes in $\mathcal{F}^{(2)}$ corresponding to a particular behavioral sequence cannot be determined a priori, any more than each mind could contain a priori representations of all the sonatas that ever were or will be composed. For a similar reason, neither the serial ordering of these nodes nor the selective attachment of an arousal source to this set of sampling nodes is determined a priori. All of these structures need to be self-organized by developmental and learning mechanisms, and we need to understand how this happens.

XI. SELF-ORGANIZATION OF PLANS: THE GOAL PARADOX

These formal problems concerning the self-organization of avalanches also arise in many types of goal-oriented behavior. Consider maze learning (Figure 6.7) as an idealization of the many situations in which one learns a succession of choice points leading to a goal, such as going from home to the store or from one's office to the cafeteria. In Figure 6.7, one leaves the filled-in start box and is rewarded with food in the vertically hatched goal box. After learning has occurred, every errorless transit from start box to goal box requires the same sequence of turns at choice points in the maze. In some sense, therefore, our memory traces can encode the correct order at the choice points. Moreover, the goal box always occurs *last* on every learning trial, whether or not errors occur. During goal-directed performance, by contrast, the *first* thing that we think of is the goal, which somehow activates a plan that generates a correct series of actions. How is it possible for the goal event to always occur last on learning trials and for memory to encode the correct perfor-

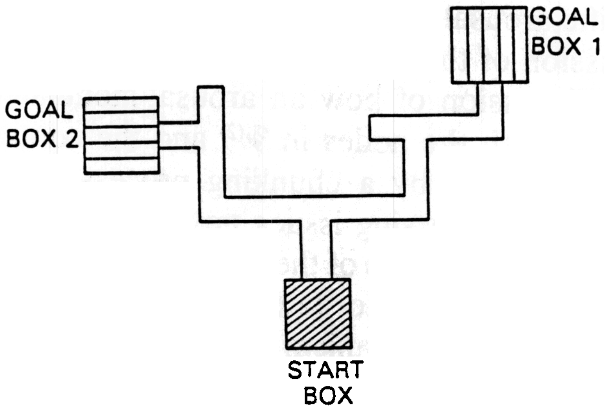

Figure 6.7 The Goal Paradox: Correct performance from start box to a goal box is always order-preserving. On every learning trial, the goal always occurs last. How can the correct order be learned despite the fact that an internal representation of the goal is activated first on performance trials and organizes the correct ordering of acts that preceded the goal on every learning trial?

mance order, and yet for an internal representation of the goal to activate commands corresponding to events that preceded the goal on every learning trial? This is the *goal paradox* (Grossberg, 1978c). Actually, the goal paradox seems paradoxical only if one takes seriously the manner in which a learning subject's information unfolds in real time and vigorously resists the temptation to beg the question by appealing to a homunculus.

To overcome the goal paradox, we need to understand how the individual sensory–motor coordinations leading to a correct turn at each choice point can be self-organized. This problem is formally analogous to the problem of self-organizing the individual nodes $\mathcal{F}^{(2)}$ of a context-modulated avalanche and of associating each of these nodes with a motor command across $\mathcal{F}^{(1)}$. We also need to understand how a plan can serially organize individual choices into a correctly ordered sequence of choices. This latter problem needs to be broken into at least two successive stages.

First, the internal representation corresponding to the plan is determined by the internal representations of the individual choices themselves, since during a correct transit of the maze on a learning trial, these are the relevant data that are experienced. The dependence of the plan on its defining sequence of events is schematized by the upward directed arrow in Figure 6.7a. The process whereby a planning node is selected is a form of code development, or chunking. After the plan is self-organized by its defining sequence of events, the plan in turn learns which internal representations have activated it, so that it can selectively activate only these representations on performance trials (Figure 6.8b). A feedforward, or bottom-up, process of coding (chunking) thus sets the stage for a feedback, or top-down, process of pattern learning (expectancy learning). In its role as a source of contextual feedback signals, the plan is analogous to the nonspecific arousal node $v_1^{(3)}$ in the context-modulated avalanche.

This discussion of the goal paradox emphasizes an issue that was implicit in my discussion of how an arousal-modulated avalanche can be self-organized. Both the nodes in $\mathcal{F}^{(2)}$ and the context-sensitive node in $\mathcal{F}^{(3)}$ are self-organized by a chunking process. Given this conclusion, some specialized processing issues now come into view. What mechanism maintains the activities of the $\mathcal{F}^{(2)}$ nodes long enough for the simultaneously active $\mathcal{F}^{(2)}$ nodes to determine and be determined by the $\mathcal{F}^{(3)}$ planning node? What mechanism enables the planning node to elicit the correct performance order from the $\mathcal{F}^{(2)}$ nodes? Are learned associations between the $\mathcal{F}^{(2)}$ nodes necessary, as is suggested by the links $v_i^{(2)} \rightarrow v_{i+1}^{(2)}$ of a context-modulated avalanche, or can the learned signals from $\mathcal{F}^{(3)}$ to $\mathcal{F}^{(2)}$ encode order information by themselves?

Once we explicitly recognize that the sequence of active $\mathcal{F}^{(2)}$ nodes determines its own $\mathcal{F}^{(3)}$ node, we must also face the following problem.

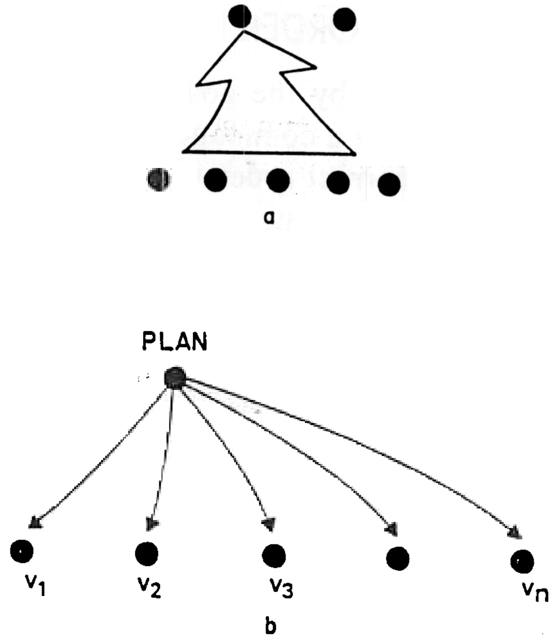

Figure 6.8 The feedback loop between chunking and planned temporal ordering: (a) Activity patterns across the item representations choose their planning nodes via a process of feedforward (bottom-up) code learning (chunking). (b) The chosen planning nodes can learn to activate the item representations which chose them, and the order with which items are activated, by a process of feedback (top-down) associative pattern learning (expectancy learning).

Every subsequence of an event sequence is a perfectly good event sequence in its own right. By the principle of sufficient reason, each subsequence may therefore encode its own node. Not every subsequence may be able to encode a planning node with equal ease. Nonetheless, the single nonspecific arousal source $v_1^{(3)}$ needs to be replaced by a field $\mathcal{F}^{(3)}$ of command nodes which are activated by prescribed subsequences across $\mathcal{F}^{(2)}$. Every node in $\mathcal{F}^{(3)}$ can in turn sample the activity pattern across $\mathcal{F}^{(2)}$ while it is active. As performance proceeds, the event sequences represented across $\mathcal{F}^{(2)}$ and the planning nodes activated across the field $\mathcal{F}^{(3)}$ continually change and mutually influence one another. A more advanced version of the same problem arises when, in addition to feedback exchanges within sensory and motor modalities, intermodality reciprocal exchanges control the unfolding of recognition and recall through time (Grossberg, 1978e).

The deepest issues relating to these feedback exchanges concern their stability and self-consistency. What prevents every new event in $\mathcal{F}^{(2)}$ from destroying the encoding of past event sequences in $\mathcal{F}^{(3)}$? What enables the total output from $\mathcal{F}^{(3)}$ to define a more global and predictive context than could any individual planning node?

XII. TEMPORAL ORDER INFORMATION IN LTM

The design issues raised by the goal paradox will be approached in stages. First I consider how a command node can learn to read out item representations in their correct order without using learned associations between the item representations themselves. I also consider how ordered associations among item representations can be learned. Both of these examples are applications of serial learning properties that were first analyzed in Grossberg (1969f) and generalized in Grossberg and Pepe (1970, 1971). The temporal order information in both examples is learned using the associative learning Equations 6 and 7. Both types of temporal order information can be learned simultaneously in examples wherein pathways within $\mathcal{F}^{(2)}$ and between $\mathcal{F}^{(2)}$ and $\mathcal{F}^{(3)}$ are conditionable. Both examples overcome the instability problem of Estes's (1972) model of temporal order in LTM, and neither example requires a serial buffer to learn its temporal order information. Murdock (1979) has studied serial learning properties using a related approach but one that is weakened by the conceptual difficulties of linear system theory models (Section II). In particular, the Murdock approach has not yet been able to explain the bowed and skewed cumulative error curve in serial verbal learning.

The problem of how a command node learns to read out an STM pattern that encodes temporal order information across item representations has two different versions, depending on whether the command node is excited before the first list item is presented or after the last list item is presented. The former problem will be considered now, but the latter problem cannot be considered until Section XXXIV, since it requires a prior analysis of competitive STM interactions for its solution (Section XVII).

XIII. READ-OUT AND SELF-INHIBITION OF ORDERED STM TRACES

Figure 6.9a depicts the desired outcome of learning. The LTM traces z_{1i} from the command node $v_1^{(3)}$ to the item representations $v_i^{(2)}$ satisfy the chain of inequalities

$$z_{11} > z_{12} > z_{13} > \cdot \cdot \cdot > z_{1m} \tag{24}$$

due to the fact that the list of items $r_1, r_2, \ldots, r_m$ was previously presented to $\mathcal{F}^{(2)}$. Consequently, when a performance signal from $v_1^{(3)}$ is gated by these LTM traces, an STM pattern across $\mathcal{F}^{(2)}$ is generated such that

$$x_1^{(2)} > x_2^{(2)} > x_3^{(2)} > \cdot \cdot \cdot > x_m^{(2)}. \tag{25}$$

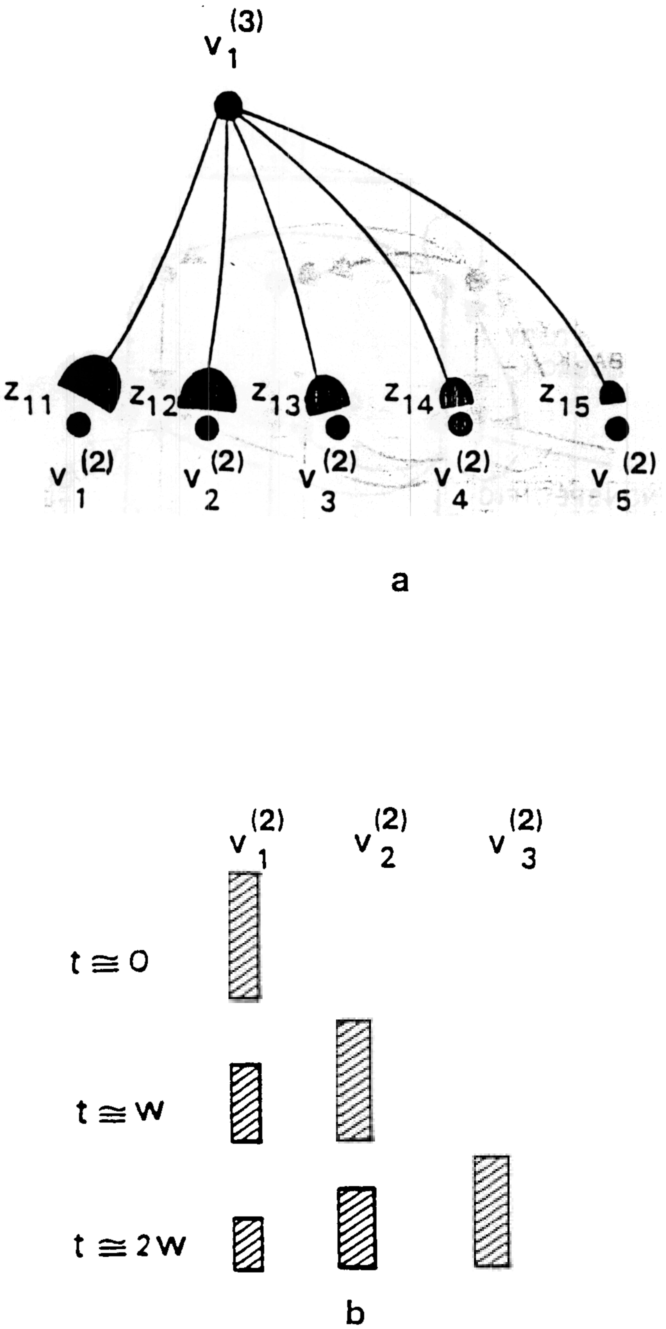

Figure 6.9 Simultaneous encoding of context and temporal order by top-down STM–LTM order reversal: (a) The context node $v_1^{(3)}$ reads out a primacy gradient across the item representations of $\mathcal{F}^{(2)}$. (b) The context node $v_1^{(3)}$ can learn a primacy gradient in LTM by multiplicatively sampling and additively storing a temporal series of STM recency gradients across $\mathcal{F}^{(2)}$.

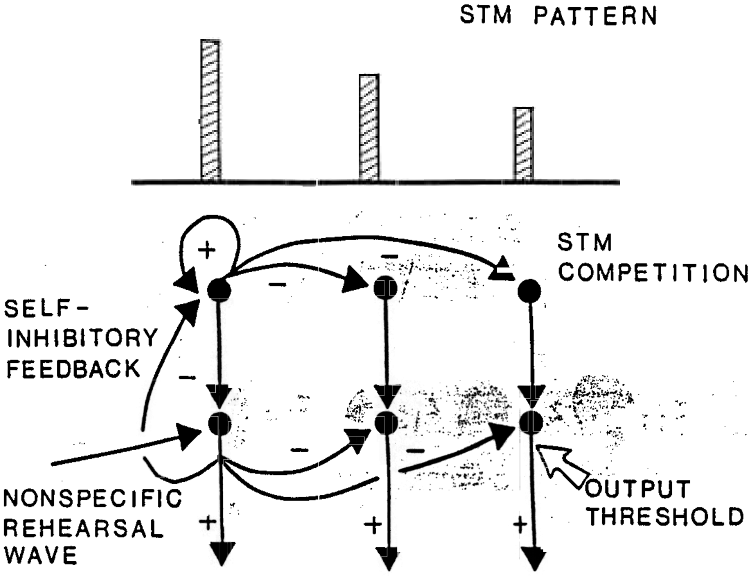

Figure 6.10 A reaction time rule translates larger STM activities into faster output onsets. Output-contingent STM self-inhibition prevents item perseveration.

A reaction time rule such as Equation 23 initiates an output signal faster from a node with a large STM activity than from a node with a small STM activity. The chain of STM inequalities (Equation 25) can thus be translated into the correct order or performance using such a reaction time rule if the following problem of perseveration can be prevented. After the first item r_1 is performed, the output signal from $v_1^{(2)}$ must shut off to prevent a sustained output signal from interfering with the performance of later items. A specific inhibitory feedback pathway thus inhibits $x_1^{(2)}$ after a signal is emitted from $v_1^{(2)}$ (Figure 6.10). The same perseveration problem then faces the remaining active nodes $v_2^{(2)}, v_3^{(2)}, \ldots, v_m^{(2)}$. Hence every output pathway from $\mathcal{F}^{(2)}$ can activate a specific inhibitory feedback pathway whose activation can self-inhibit the corresponding STM trace (Grossberg, 1978a, 1978e; Rumelhart & Norman, 1982). With this performance mechanism in hand, we now consider the more difficult problem of how the chain of LTM inequalities (Equation 24) can be learned during presentation of a list of items $r_1, r_2, r_3, \ldots, r_m$.

XIV. THE PROBLEM OF STM–LTM ORDER REVERSAL

The following example illustrates the problem in its most severe form. The STM properties that I now consider will, however, have to be generalized in Section XXXIV. Suppose that each node $v_i^{(2)}$ is excited by a fixed amount when the ith list item r_i is presented. Suppose also that as time goes on, the STM trace $x_i^{(2)}$ gets smaller due either to internodal competition or passive trace decay. Which of the two decay mechanisms is used does not affect the basic result, although different mechanisms will cause testable secondary differences in the order information to be encoded in LTM. Whichever decay mechanism is used, in response to serially presented lists, the last item to have occurred always has the largest STM trace. In other words, a recency effect exists at each time in STM (Figure 6.9b). Given this property, how is the chain of LTM inequalities learned? In other words, how does a sequence of recency gradients in STM get translated into a primacy gradient in LTM? I call this issue the "STM–LTM order reversal problem" (Grossberg, 1978e).

The same problem arises during serial verbal learning but in a manner that disguises its relevance to planned serial behavior. In this task, the generalization gradients of errors at each list position have the qualitative form depicted in Figure 6.11. A gradient of anticipatory (forward) errors occurs at the beginning of the list, a two-sided gradient of anticipatory and perseverative (backward) errors near the middle of the list, and a gradient of perseverative errors at the end of the list (Osgood, 1953). I suggest that the gradient of anticipatory errors at the beginning of the list is learned in the same way as a primacy gradient in LTM. I have shown (Grossberg, 1969f) that the same associative laws also generate the other position-sensitive error gradients. Thus a command node that is activated after the entire list is presented encodes a recency gradient in LTM rather than the primacy gradient that is encoded by a command node activated before (or when) the first list item is presented. The same laws also provide an explanation of why the curve of cumulative errors versus list position is bowed and skewed toward the end of the list, and of why associations at

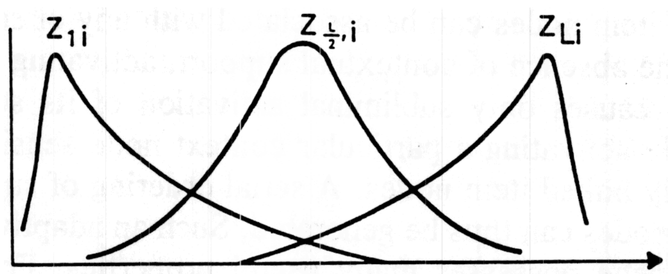

Figure 6.11 Each sampling node v_j learns a different LTM pattern $z_j = (z_{j1}, z_{j2}, \ldots, z_{jn})$ if it samples at different times. In a list of length $n = L$ whose intertrial interval is sufficiently long, a node that starts sampling at the list beginning ($j \cong 1$) learns a primacy gradient in LTM. At the list end ($j \cong L$), a recency gradient in LTM is learned. Near the list middle ($j \cong L/2$), a two-sided LTM gradient is learned. When STM probes read different LTM patterns z_j into STM, the different patterns generate different error gradients due to the action of internal noise, the simultaneous read-out by other probes, and the STM competition that acts after LTM read-out.

the beginning of the list are often, but not always, learned faster than associations at the end of the list (Grossberg & Pepe, 1970, 1971).

From the perspective of planned serial behavior, these results show how the activation of a command node at different times during list presentation causes the node to encode totally different patterns of order information in LTM. Thus the learning of order information is highly context-sensitive. If a command node is activated by a prescribed list subsequence via $\mathcal{F}^{(2)} \rightarrow \mathcal{F}^{(3)}$ signals that subserve chunking and recognition, then this subsequence constrains the order information that its command node will encode by determining the time at which the command node is activated. Moreover, this context sensitivity is not just a matter of determining which item representations will be sampled, as the issue of STM–LTM order reversal clearly shows.

An important conclusion of this analysis is that the same sort of context-sensitive LTM gradients are learned on a single trial regardless of whether command nodes sample item representations at different times or if the item representations sample each other through time. Although the order information that is encoded by the sampling nodes is the same, the two situations are otherwise wholly distinct. In the former case, list subsequences are the functional units that control learned performance, and many lists can be learned and performed over the same set of item representations. In the latter case, individual list items are the functional units that control learned performance, and once a given chain of associations is learned among the item representations, it will interfere with the learning of any other list ordering that is built up from the same item representations (Dixon & Horton, 1968; Lenneberg, 1967; Murdock, 1974).

A third option is also available. It arises by considering a context-modulated avalanche whose serial ordering and context nodes are both self-organized by associative processes (Figure 6.12). In such a network, each of the item nodes can be associated with any of several other item nodes. In the absence of contextual support, activating any one of these item nodes causes only subliminal activation of its set of target item nodes, while activating a particular context node sensitizes a subset of associatively linked item nodes. A serial ordering of supraliminally activated item nodes can thus be generated. Such an adaptive context-modulated avalanche possesses many useful properties. For example, item nodes are no longer bound to each other via rigid associative chains. Hence, a given item can activate different items in different contexts. The inhibition of a given context node can rapidly prevent the continued performance of the item ordering that it controls, while the activation of different context nodes can rapidly instate a new performance ordering among the same items or different items.

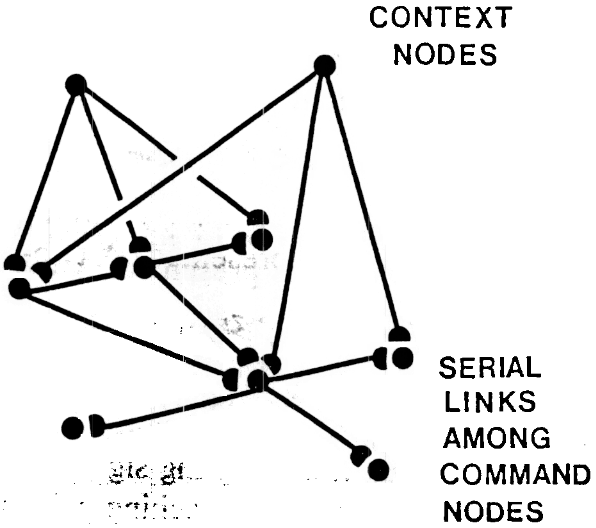

Figure 6.12 In an adaptive context-modulated avalanche, each command node can be associated with a set of command nodes that it subliminally activates. Learned top-down signals from a context node can also sensitize a set of command nodes. The convergent effects of top-down and internodal signals causes the supraliminal activation of command nodes in a prescribed serial order. Different context nodes can generate different serial orders.

In Figure 6.12, the item nodes are called command nodes. This change of terminology is intended to emphasize the fact that in order for this design to be useful, the items must represent chunks on a rather high level of network processing. The number of transitions from each command node to its successors must be reasonably small in order to achieve the type of unambiguous serial ordering that the context chunk is supposed to guarantee. The sequence chunks within the masking field discussed in Section XXXVIII are prime candidates for command nodes in an adaptive context-mediated avalanche. The ability of top-down and serial associative signals to activate ordered STM traces supraliminally without also unselectively activating a broad field of STM traces is facilitated by balancing these excitatory associative signals against inhibitory signals, notably inhibitory masking signals (see Section XXXVIII; Grossberg, 1978e, Sections 41–46).

XV. SERIAL LEARNING

This section indicates how the context-sensitive LTM gradients in Figure 6.11 are learned. Why the same rules imply that the cumulative error curve of serial learning is bowed and skewed is reviewed from a recent perspective in Grossberg 1982a. First, I consider how a primacy gradient

(Equation 24) is encoded by the LTM traces $(z_{11}, z_{12}, \ldots, z_{1m})$ of a nodc $v_1^{(3)}$ that is first activated before, or when, the first list item is presented. I then show how a recency gradient

$$z_{n1} < z_{n2} < \cdots < z_{nm}$$

is encoded by the LTM traces $(z_{n1}, z_{n2}, \ldots, z_{nm})$ of a node $v_n^{(3)}$ that is first activated after the whole list is presented. A two-sided gradient

$$z_{k1} < z_{k2} < \cdots < z_{kr} > z_{k,r+1} > \cdots > z_{km} \tag{27}$$

encoded by a node $v_k^{(2)}$ that is activated during the midst of the list presentation can then be understood as a combination of these effects.

Let node $v_1^{(3)}$ start sending out a sampling signal E_1 at about the time that r_1 is being presented. After rapidly reaching peak size, the signal E_1 gradually decays through time with its STM trace $x_1^{(3)}$ as future list items $r_2, r_3, \ldots$ are presented. Thus E_1 is largest when STM trace $x_1^{(2)}$ is maximal, smaller when both traces $x_1^{(2)}$ and $x_2^{(2)}$ are active, smaller still when traces $x_1^{(2)}$, $x_2^{(2)}$, and $x_3^{(2)}$ are active, and so on. Consequently, the product $E_1x_1^{(2)}$ in row 1 of Figure 6.9b exceeds the product $E_1x_2^{(2)}$ in row 2 of the figure, which in turn exceeds the product $E_1x_3^{(2)}$ in row 3, and so on. Due to the slow decay of each LTM trace z_{1i} on each learning trial, z_{11} adds up the products $E_1x_1^{(2)}$ in successive rows of column 1, z_{12} adds up the products $E_1x_2^{(2)}$ in successive rows of column 2, and so on. An LTM primacy gradient (Equation 24) is thus generated due to the way in which E_1 samples the successive STM recency gradients, and to the fact that the LTM traces z_{1i} add up the sampled STM gradients $E_1x_i^{(2)}$.

By contrast, the sampling signal E_n emitted by node $v_n^{(3)}$ samples a different set of STM gradients because $v_n^{(3)}$ starts to sample only after all the item representations $v_1^{(2)}, v_2^{(2)}, \ldots, v_m^{(2)}$ have already been activated on a given learning trial. Consequently, when the sampling signal E_n does turn on, it sees the already active STM recency gradient

$$x_1^{(2)} < x_2^{(2)} < \cdots < x_m^{(2)}$$

of the entire list. Moreover, the ordering (Equation 28) persists for a while because no new items are presented until the next learning trial. Thus signal E_n samples an STM recency gradient at every time. When all sampled recency gradients are added up through time, they generate a recency gradient (Equation 26) in the LTM traces of $v_n^{(3)}$. In summary, command nodes that are activated at the beginning, middle, or end of a list encode different LTM gradients because they multiplicatively sample STM patterns at different times and summate these products through time.

XVI. RHYTHM GENERATORS AND REHEARSAL WAVES

The previous discussion forces two refinements in our ideas about how nonspecific arousal is regulated. In a context-modulated avalanche, the nonspecific arousal node $v_1^{(3)}$ both selects the set of nodes $v_i^{(2)}$ that it will control and continuously modulates performance velocity across these nodes. A command node that reads out temporal order information as in Figure 6.9a can no longer fulfill both roles. Increasing or decreasing the command node's activity while it reads out its LTM pattern proportionally amplifies the STM of all its item representations. Arbitrary performance rhythms are no longer attainable, because the relative reaction times of individual item representations are constrained by the pattern of STM order information. Nor is a sustained but continuously modulated supraliminal read-out from the command node permissible, because item representations that were already performed could then be reexcited, leading to a serious perseveration problem.

Thus, if a nonspecific arousal source dedicated to rhythmic control is desired, it must be distinguished from the planning nodes. Only then can order information and rhythm information remain decoupled. The reader should not confuse this idea of rhythm with the performance timing that occurs when item representations are read out as fast as possible (Sternberg, Monsell, Knoll, & Wright, 1978; Sternberg, Wright, Knoll, & Monsell, 1980). Properties of such a performance can, in fact, be inferred from the mechanism for read-out of temporal order information per se (Section XLVII).

Another type of nonspecific arousal is also needed. If read-out of LTM order information is achieved by activating the item representations across $\mathcal{F}^{(2)}$, what prevents these item representations from being uncontrollably rehearsed, and thereby self-inhibited, while the list is being presented? To prevent this from happening, it is necessary to distinguish between STM activation of an item representation and output signal generation by an active item representation. This distinction is mechanized by assuming the existence of a nonspecific rehearsal wave capable of shunting the output pathways of the item representations. When the rehearsal wave is off, the item representations can blithely reverberate their order information in STM without generating self-destructive inhibitory feedback. Only when the rehearsal wave turns on does the read-out of order information begin.

The distinction between STM storage and rehearsal has major implications for which planning nodes in $\mathcal{F}^{(3)}$ will be activated and what they will learn. This is due to two facts working together: The rehearsal wave can determine which item subsequences will be active at any moment by rehearsing, and thereby inhibiting, one group of item representations before the next group of items is presented. Each active subsequence of item representations can in turn chunk its own planning node. The rehearsal wave thus mediates a subtle interaction between the item sequences that occur and the chunks that form to control future performance (Section XXXVII).

XVII. SHUNTING COMPETITIVE DYNAMICS IN PATTERN PROCESSING AND STM: AUTOMATIC SELF-TUNING BY PARALLEL INTERACTIONS

This analysis of associative mechanisms suggests that the unit of LTM is a spatial pattern. This result raises the question of how cellular tissues can accurately register input patterns in STM so that LTM mechanisms may encode them. This is a critical issue in cells because the range over which cell potentials, or STM traces, can fluctuate is finite and often narrow compared to the range over which cellular inputs can fluctuate. What prevents cells from routinely turning on all their excitable sites in response to intense input patterns, thereby becoming desensitized by saturation before they can even register the patterns to be learned? Furthermore, if small input patterns are chosen to avoid saturation, what prevents the internal noise of the cells from distorting pattern registration? This *noise–saturation dilemma* shows that cells are caught between two potentially devastating extremes. How do they achieve a ''golden mean'' of sensitivity that balances between these extremes?

I have shown (Grossberg, 1973) that mass action competitive networks can automatically retune their sensitivity as inputs fluctuate to register input differences without being desensitized by either noise or saturation. In a neural context, these systems are called shunting on-center off-surround networks. The shunting, or mass action, dynamics are obeyed by the familiar membrane equations of neurophysiology; the automatic retuning is due to automatic gain control by the inhibitory signals.

The fixed operating range of cells should not be viewed as an unmitigated disadvantage. By fixing their operating range once and for all, cells can also define fixed output thresholds and other decision criteria with respect to this operating range. By maintaining sensitivity within this operating range despite fluctuations in total input load, cells can achieve an impressive parallel processing capability. Even if parallel input sources to the cells switch on or off unpredictably through time, thereby changing the total input to each cell, the automatic gain control mechanism can recalibrate the operating level of total STM activity to bring it into the range of the cells' fixed decision criteria. Additive models, by contrast, do not have this capability. These properties are mathematically described in Grossberg (1983, Sections 21–23).

Because the need to accurately register input patterns by cells is ubiquitous in the nervous system, competitive interactions are found at all levels of neural interaction and of my models thereof. A great deal of what is called "information processing" in other approaches to intelligence reduces in the present approach to a study of how to design a competitive, or close-to-competitive, network to carry out a particular class of computations. Several types of specialized competitive networks will be needed. As I mentioned in Section I, the class of competitive systems includes examples which exhibit arbitrary dynamical behavior. Computer simulations that yield an interesting phenomenon without attempting to characterize which competitive parameters control the phenomenon teach us very little, because a small adjustment of parameters could, in principle, generate the opposite phenomenon. To quantitatively classify the parameters that control biologically important competitive networks is therefore a major problem for theorists of mind. Grossberg (1981a, Sections 10–27) and Cohen and Grossberg (1983) review some results of this ongoing classification.

XVIII. CHOICE, CONTRAST ENHANCEMENT, LIMITED STM CAPACITY, AND QUENCHING THRESHOLD

Some of the properties that I use can be illustrated by the simplest type of competitive feedback network:

$$\frac{\mathrm{d}}{\mathrm{dt}} x_i = -Ax_i + (B - x_i)[I_i + \mathrm{f}(x_i)] - x_i \left[J_i + \sum_{k \neq i} \mathrm{f}(x_k)\right]$$

where $i = 1, 2, \ldots, n$. In Equation 29, term $-Ax_i$ describes the passive decay of the STM trace x_i at rate $-A$. The excitatory term $(B - x_i)[I_i + \mathrm{f}(x_i)]$ describes how an excitatory input I_i and an excitatory feedback signal $\mathrm{f}(x_i)$ from v_i to itself excites by mass action the unexcited sites $(B - x_i)$ of the total number of sites B at each node v_i. The inhibitory term $-x_i[J_i + \Sigma_{k \neq i} \mathrm{f}(x_k)]$ describes how the inhibitory input J_i and the inhibitory, or competitive, feedback signals $\mathrm{f}(x_k)$ from all v_k, $k \neq i$, turn off the x_i excited sites of v_i by mass action.

Equation 29 strips away all extraneous factors to focus on the following issue. How does the choice of the feedback signal function $\mathrm{f}(w)$ influence the transformation and storage of input patterns in STM? To discuss this network's positive feedback loops. This fact represents a serious challenge to linear feedback models (Grossberg, 1978d).

A faster-than-linear signal function can tell the difference between small and large inputs by amplifying and storing only sufficiently large activities. Such a signal function amplifies the larger activities so much more than the smaller activities that it makes a choice: Only the largest initial activity is stored in STM. A sigmoid signal function can also suppress noise, although it does so less vigorously than a faster-than-linear signal function. Consequently, activities less than a criterion level, or quenching threshold (QT), are suppressed, whereas the pattern of activities that exceeds the QT is contrast-enhanced before being stored in STM.

Any network that possesses a QT can be *tuned.* By increasing or decreasing the QT, the criteria of which activities represent functional signals—and hence should be processed and stored in STM—and of which activities represent functional noise—and hence should be suppressed—can be flexibly modified through time. An increase in the QT can cause all but the largest activities to be quenched. Thus the network can behave like a choice machine if its storage criteria are made sufficiently strict. A sudden decrease in the QT can cause all recently presented patterns to be stored. If a novel or unexpected event suddenly decreases the QT, all relevant data can be stored in STM until the cause of the unexpected event is learned (Grossberg, 1975, 1982b). It cannot be overemphasized that the existence of the QT and its desirable tuning properties all follow from the use of a nonlinear signal function.

To illustrate the QT concept concretely, consider a sigmoid signal function f(w) that is faster than linear for $0 \leq w \leq x^{(1)}$ and linear for $x^{(1)} \leq w \leq B$. The slower-than-linear part of f(w) does not affect network dynamics because each $x_i \leq B$ by Equation 29. More precisely, let $f(w) = Cwg(w)$, where $C \geq 0$, g(w) is increasing for $0 \leq w \leq x^{(1)}$, and $g(w) = 1$ for $x^{(1)} \leq w \leq B$. Grossberg (1973, pp. 355–359) has demonstrated that the QT of this network is

$$\frac{x^{(1)}}{B - AC^{-1}}$$

By this equation, the QT is not the "manifest" threshold of f(w), which occurs in the range where g(w) is increasing. Instead, the QT depends on the transition activity $x^{(1)}$ at which the signal function becomes linear, the slope C of the signal function, the number of excitable sites B, and the STM decay rate A. Thus all the parameters of the network influence the size of the QT. By Equation 30, an increase in C causes a decrease in the QT. In other words, increasing a shunting signal C that nonspecifically gates all the network's feedback pathways facilitates STM storage.

question, I assume that inputs $(I_1, I_2, \ldots, I_n, J_1, J_2, \ldots, J_n)$ are delivered before time $t = 0$ and switch off at time $t = 0$ after having instated an initial pattern $x(0) = (x_1(0), x_2(0), \ldots, x_n(0))$ in the network's STM traces. Our task is to understand how the choice of f(w) influences the transformation of $x(0)$ into the stored pattern $x(\infty) = (x_1(\infty), x_2(\infty), \ldots, x_n(\infty))$ as time increases.

Figure 6.13 shows that different choices of f(w) generate markedly different storage modes. The function $g(w) = x^{-1}f(w)$ is also graphed in Figure 6.13 because the property that determines the type of storage is whether g(w) is an increasing, constant, or decreasing function at prescribed values of the activity w. For example, as in the four rows of Figure 6.13, a linear $f(w) = aw$ generates a constant $g(w) = a$; a slower-than-linear $f(w) = aw(b + w)^{-1}$ generates a decreasing $g(w) = a(b + w)^{-1}$; a faster-than-linear $f(w) = aw^n$, $n > 1$, generates an increasing $g(w) = aw^{n-1}$; and a sigmoid $f(w) = aw^2(b + w^2)^{-1}$ generates a concave $g(w) = aw(b + w^2)^{-1}$. Both linear and slower-than-linear signal functions amplify noise. Even tiny activities are bootstrapped into large activities by the

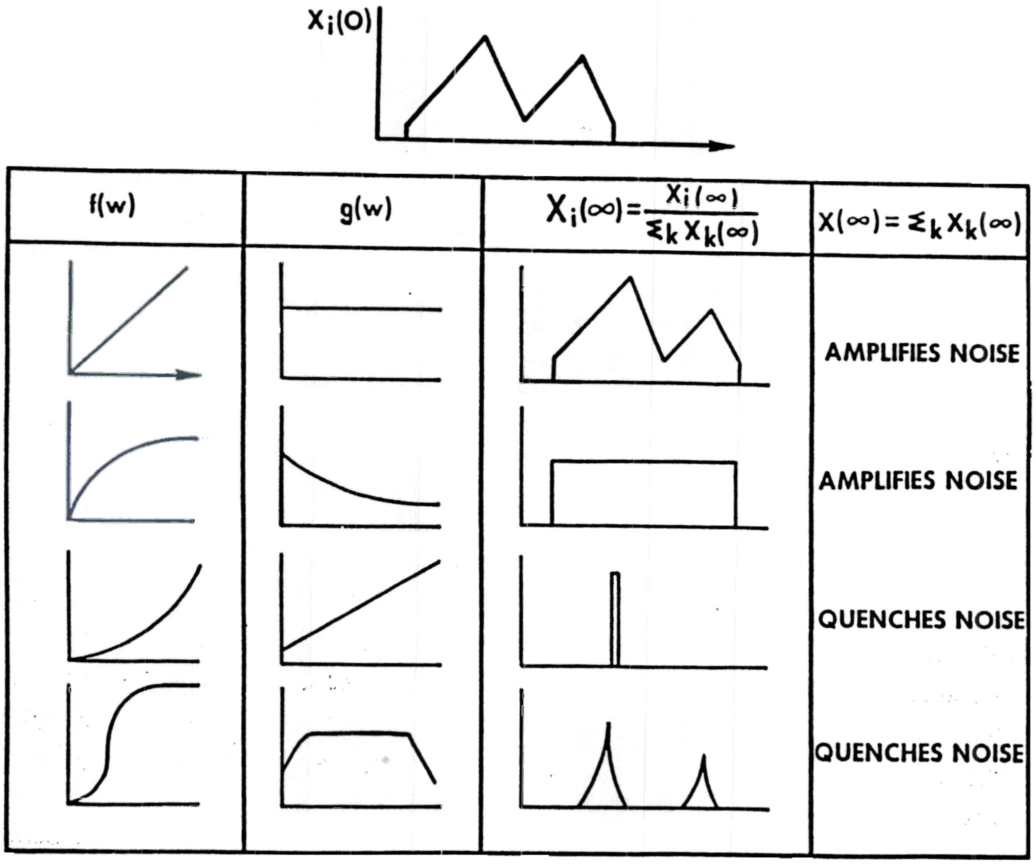
$X_i(0)$
f(w)
g(w)
$X_i(\infty) = \frac{X_i(\infty)}{\Sigma_k X_k(\infty)}$
$X(\infty) = \Sigma_k X_k(\infty)$
AMPLIFIES NOISE
AMPLIFIES NOISE
QUENCHES NOISE
QUENCHES NOISE

Figure 6.13 Influence of signal function f(w) on input pattern transformation and STM storage.

Another property of STM in a competitive network is its limited capacity. This property follows from the network's tendency to conserve, or normalize, the total suprathreshold activity that it can store in STM. Consequently, an increase in one STM trace forces a decrease in other STM traces. As soon as one of these diminished traces becomes smaller than the QT, it is suppressed.

A full understanding of the normalization concept, no less than the QT concept, requires a mathematical study of relevant examples. The case wherein f(w) is faster-than-linear illustrates normalization in its simplest form. Let $x = \Sigma_{i=1}^{n} x_i$ be the total STM activity, and let $F = \Sigma_{i=1}^{n} \mathrm{f}(x_i)$ be the total feedback signal. Summing over the index i in Equation 29 yields the equation

$$\frac{\mathrm{d}}{\mathrm{dt}}x = -Ax + (B - x)F. \tag{31}$$

To solve for possible equilibrium activities $x(\infty)$ of $x(t)$, let $(\mathrm{d}/\mathrm{dt})x = 0$ in Equation 31. Then

$$\frac{Ax}{B - x} = F.$$

Since a network with a faster-than-linear feedback signal makes a choice, only one STM trace $x_i(t)$ remains positive at $t \to \infty$. Hence only one summand in F remains positive as $t \to \infty$, and its $x_i(t)$ value approaches $x(t)$. Consequently, Equation 32 can be rewritten as

$$\frac{Ax}{B - x} = \mathrm{f}(x).$$

Equation 33 is independent of the number of active nodes. Hence the total STM activity is independent of the number of active nodes.

XIX. LIMITED CAPACITY WITHOUT A BUFFER: AUTOMATICITY VERSUS COMPETITION

The formal properties of the previous section are reflected in many types of data. A fixed capacity buffer is often posited to explain the limited capacity of STM (Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981). Such a buffer is often implicitly or explicitly endowed with a serial ordering of buffer positions to explain free recall data. Buffer models do not, however, explain how items can be read in to one buffer position and still move their representations along buffer positions in such a way that *every* item can be performed from each buffer position, as is required for the buffer idea to meet free recall data. The buffer concept also tacitly implies that the entire hierarchy of codes that is derivable from item representations can also be shifted around as individual item codes move in the buffer.

The normalization property provides a dynamical explanation of why STM has a limited capacity without using a serial buffer. In the special case that new item representations get successively excited by a list of inputs, the normalization property implies that other item representations must lose activity. As soon as one of the activities falls below the QT, it drops out of STM. No motion of item representations through a buffer is required. Hence, no grueling problems of shift-invariant read-in and read-out need to be solved.

In this view of the limited capacity of STM, it is important to know which item representations are mutually inhibitory. Equation 29 represents the atypical situation in which each item representation can inhibit all other item representations with equal ease via the inhibitory terms $\Sigma k \neq i$ f(x_k). More generally, an equation of the form

$$\frac{\mathrm{d}}{\mathrm{dt}} x_i = -A_i x_i + (B_i - x_i)[I_i + \sum_{k=1}^{n} \mathrm{f}_k(x_k) C_{ki}]$$
$$- (x_i + D_i)[J_i + \sum_{k=1}^{n} \mathrm{g}_k(x_k) E_{ki}]$$

holds, $i = 1, 2, \ldots, n$, in which the excitatory signal $\mathrm{f}_k(x_k)$ from v_k excites v_i with a strength $\mathrm{f}_k(x_k)C_{ki}$, whereas the inhibitory signal $\mathrm{g}_k(x_k)$ from v_k inhibits v_i with a strength $\mathrm{g}_k(x_k)E_{ki}$. If the inhibitory coefficients E_{ki} decrease with the network distance between v_k and v_i, then total STM activity can progressively build up as more items are presented until the density of active nodes causes every new input to be partly inhibited by a previously excited node. Thus sparsely distributed items may, at least at a single network level, sometimes be instated "automatically" in STM by their inputs without incurring competitive "capacity limitations" (Norman & Bobrow, 1975; Schneider & Shiffrin, 1976, 1977). The possibility that total STM activity can build up to an asymptote plays an important part in characterizing stable rules for laying down temporal order information in STM (Section XXXIV).

"Automatic" processing can also occur in the present theory due to the influence of learned top-down expectancies, or feedback templates, on competitive matching processes (Section XXIV). The tendency to sharply differentiate automatic versus controlled types of processing has been popularized by the work of Schneider and Shiffrin (1976, 1977), who ascribe automatic properties to a parallel process and controlled properties to a serial process. This distinction creates conceptual paradoxes when it is joined to concepts about learning (e.g., Grossberg, 1978e, Section 61). Consider the serial learning of any new list of familiar items. Each familiar item is processed by a parallel process, while each unfamiliar inter-item contingency is processed by a serial process. Schneider and Shiffrin's theory thus implies that the brain somehow rapidly alternates between parallel and serial processing when the list is first presented but switches to sustained parallel processing as the list is unitized. Consider the perception of a picture whose left half contains a familiar face and whose right half contains a collection of unfamiliar features. Schneider and Shiffrin's theory implies that the brain somehow splits the visual field into a serial half and a parallel half, and that the visual field gets reintegrated into a parallel whole as the unfamiliar features are unitized.

These paradoxes arise from the confusion of serial properties with serial processes and of parallel properties with parallel processes. All the processes of the present theory are parallel processes, albeit of a hierarchically organized network. The present theory shows how both serial properties and parallel properties can be generated by these parallel processes in response to different experimental paradigms. In particular, "the auditory-to-visual *codes* and *templates* that are activated in VM [varied mapping] and CM [consistent mapping] conditions are different, but the two conditions otherwise share common mechanisms" (Grossberg, 1978e, p. 364). Some evoked potential tests of this viewpoint and explanations of other data outside the scope of the Schneider and Shiffrin theory are described elsewhere (Banquet and Grossberg, 1986; Carpenter and Grossberg, 1986b, 1986c; Grossberg, 1982d, 1984a; Grossberg and Stone, 1986a). A growing number of recent experiments also support this viewpoint (e.g., Kahneman & Chajczyk, 1983).

XX. HILL CLIMBING AND THE RICH GET RICHER

The contrast enhancement property of competitive networks manifests itself in a large body of data. A central role for both contrast enhancement and normalization has, for example, been suggested in order to search associative memory and to self-organize new perceptual and cognitive codes (Carpenter and Grossberg, 1986b; Grossberg, 1978e, 1980c, 1982b). A more direct appearance of contrast enhancement has also been suggested to exist in letter and word recognition (Grossberg, 1978e; McClelland & Rumelhart, 1981). McClelland and Rumelhart introduce a number of evocative phrases to enliven their discussion of letter recognition, such as the "rich-get-richer" effect and the "gang" effect. The former is simply a contrast enhancement effect whereby small differences in the initial activation levels of word nodes get amplified through time into larger differences. The numerical studies and intuitive arguments presented by McClelland and Rumelhart (1981) do not, however, disclose why this can sometimes happen. Figure 6.13 illustrates that a correct choice of signal function is needed for it to happen, but a correct choice of signal function is not enough to guarantee even that the network will not oscillate uncontrollably through time (Grossberg, 1978c, 1980a). The gang effect uses a reciprocal exchange of prewired feedforward filters and feedback templates between letter nodes and word nodes to complete a word representation in response to an incomplete list of letters. This type of positive feedback exchange is also susceptible to uncontrollable instabilities whose prevention has been analyzed previously (Grossberg, 1976a, 1976b, 1980c).

I prefer the use of functional words rather than shibboleths for both an important reason and frivolous reason. The important reason is that an adherence to functional words emphasizes that a single functional property is generating data in seemingly disparate paradigms. Functional words thus tend to unify the literature rather than to fragment it. The frivolous reason is that another rich-get-richer effect had already been so christened in the literature before the usage of McClelland and Rumelhart (1981), and I was the person to blame. In Grossberg (1977), I called the normative drift whereby activity can be sucked from some populations into others due to the amplified network parameters of the latter populations a rich-get-richer effect. At the time, I found this sociological interpretation both amusing and instructive. Later (Grossberg, 1978b), however, I realized that the same mechanism could elicit automatic hill climbing along a developmental gradient, and I suggested (Grossberg, 1978e) how the same hill-climbing mechanism could respond to an ambiguous stimulus by causing a spontaneous STM drift from a representation that was directly excited by the stimulus to a more complete, or normative, nearby representation which could then quickly read out its feedback template to complete the ambiguous database.

This normative mechanism was, in fact, first presented in Levine and Grossberg (1976) as a possible explanation of Gibson's (1937) line neutralization effect. Thus the same functional idea has now been used to discuss visual illusions, pattern completion and expectancy matching, developmental gradients, and even sociology. It thus needs a functional name that is neutral enough to cover all of these cases. It needs a name other than contrast enhancement because it uses a mechanism of lateral masking that is distinct from simple contrast enhancement. This type of masking will be discussed in more detail to show how item sequences can be automatically parsed in a context-sensitive fashion through time (Section XXXVIII).

XXI. INSTAR LEARNING: ADAPTIVE FILTERING AND CHUNKING

With these introductory remarks about competition in hand, I now discuss the issue of how new recognition chunks can be self-organized within the fields $\mathcal{F}^{(2)}$ and $\mathcal{F}^{(3)}$ of a context-modulated avalanche, or more generally within the command hierarchy of a goal-oriented sequence of behaviors. I first consider the minimal anatomy that is capable of chunking or code development; namely, the *instar* (Figure 6.14). As its name suggests, the instar is the network dual of an outstar. An instar is constructed from an outstar by reversing the direction of its sampling pathways. Whereas in an outstar, conditionable pathways radiate from a sampling cell to sampled cells, in an instar conditionable pathways point from sampled cells to a sampling cell. Consequently, the sampling cell of an instar is activated by a sum of LTM-gated signals from sampled cells. These signals may be large enough to activate the sampling cell and thus cause the sampled cells to be sampled. If a spatial pattern of inputs persistently excites the sampled cells, it can cause alterations in the pattern

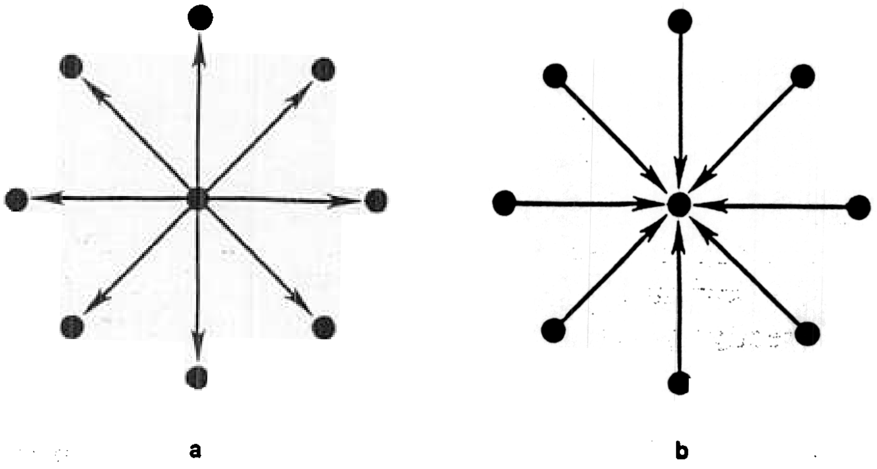

Figure 6.14 The duality between expectancy learning and chunking: (a) An *outstar* is the minimal network capable of associative pattern learning, notably expectancy learning. (b) An *instar* is the minimal network capable of code development, notably chunking. The source of an outstar excites the outstar border. The border of an instar excites the instar source. In both outstars and instars, source activation is necessary to drive LTM sampling. Since the signals from the instar border are gated by LTM traces before activating the instar source, code learning both changes the efficacy of source activation and is changed by it in an STM–LTM feedback exchange.

of LTM traces across the conditionable pathways that gate the sampled signals. These LTM changes can enable a practiced input pattern to excite the instar's sampling node with greater efficacy. The heightened activation of the sampling node measures how well the sampling source has come to represent, or chunk, the input pattern.

Although this version of chunking is too simplistic, two aspects of the problem as studied in this form have far-reaching consequences (Grossberg, 1976a, 1980c). To fix ideas, denote the sampling node by v_0 and the sampled nodes by $v_1, v_2, \ldots, v_n$. For simplicity, let the cells $v_1, v_2, \ldots, v_n$ rapidly normalize any input pattern $I_i = \theta_i I$ that they receive into STM activities $x_i = \theta_i$. Denote the signal emitted from v_i into pathway e_{i0} by $\mathrm{f}(\theta_i)$. This signal is gated by the LTM trace z_{i0} before the gated signal $\mathrm{f}(\theta_i)z_{i0}$ reaches v_0 from v_i. All of these signals are added at v_0 to get a total signal $T_0 = \Sigma_{i=1}^{n} \mathrm{f}(\theta_i)z_{i0}$. As in Equation 11, T_0 is the dot product

$$T_0 = \mathrm{f}_\theta \cdot z_0$$

of the two vectors $\mathrm{f}_\theta = (\mathrm{f}(\theta_1, \mathrm{f}(\theta_2), \ldots, \mathrm{f}(\theta_n))$ *and* $z_0 = (z_{10}, z_{20}, \ldots, z_{n0})$.

To characterize the STM response at v_0 to signal T_0, suppose for simplicity that the total activity of v_0 is normalized to 1 and that v_0 possesses a QT equal to ε. Then

$$x_0 = \begin{cases} 1, & \text{if} \quad T_0 \geq \varepsilon, \\ 0, & \text{if} \quad T_0 < \varepsilon. \end{cases} \tag{36}$$

Moreover, suppose that the LTM traces z_{i0} satisfy the equation

$$\frac{\mathrm{d}}{\mathrm{dt}} z_{i0} = [-z_{i0} + \mathrm{f}(\theta_i)]x_0. \tag{37}$$

Equation 37 is a special case of Equation 7 in which no learning occurs unless T_0 succeeds in exciting x_0. When learning does occur, z_{i0} is attracted towards $\mathrm{f}(\theta_i)$, $i = 1, 2, \ldots, n$.

Under these conditions, it is easily shown that as input trials proceed, the vector $z_0(t)$ of LTM traces is monotonically attracted to the vector f_θ of signal pattern weights. In other words, $z_0(t)$ becomes parallel to f_θ as practice proceeds. This trend tends to maximize the signal $T_0(t) = \mathrm{f}_\theta \cdot z_0(t)$ as trials proceed because the dot product is maximized when vectors of fixed length are made parallel to each order. As $T_0(t)$ grows, its ability to activate v_0 by exceeding its QT also increases. Speaking intuitively, v_0 "codes" θ due to learning trials, or the adaptive filter T_0 is "tuned" by experience.

Unfortunately, this example is deficient in several ways. For one, there is no nontrivial coding: v_0 can become more sensitive to a single pattern θ, but no one node can differentially encode a large number of patterns into more than two categories. Clearly, more sampling nodes are needed. This can be accomplished as follows.

XXII. SPATIAL GRADIENTS, STIMULUS GENERALIZATION, AND CATEGORICAL PERCEPTION

Let the nodes $v_1, v_2, \ldots, v_n$ be replaced by a field $\mathcal{F}^{(1)}$ of nodes, and let v_0 be replaced by a field $\mathcal{F}^{(2)}$ of nodes. Each pattern across $\mathcal{F}^{(1)}$ can now send a positive signal to many nodes of $\mathcal{F}^{(2)}$. How is an increasingly selective response across $\mathcal{F}^{(2)}$ to be achieved as sequences of input patterns perturb $\mathcal{F}^{(1)}$?

Both networks $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ include competitive interactions to solve the noise–saturation dilemma. The easiest way to achieve learned selectivity is thus to design $\mathcal{F}^{(2)}$ as a sharply tuned competitive feedback network that chooses its maximal input for STM storage, quenches all other inputs, and normalizes its total STM activity. By analogy with the previous example, let the total input to $v_j^{(2)}$ in $\mathcal{F}^{(2)}$ equal

$$T_j = \mathrm{f}_\theta \cdot z_j,$$

let

$$x_j^{(2)} = \begin{cases} 1, & \text{if } \; T_j \geq \max[\varepsilon, T_k : k \neq j] \\ 0, & \text{if } \; T_j < \max[\varepsilon, T_k : k \neq j], \end{cases}$$

and let

$$\frac{\mathrm{d}}{\mathrm{dt}} z_{ij} = (-z_{ij} + f(\theta_i))x_j^{(2)}.$$

Now let a sequence of spatial patterns perturb $\mathcal{F}^{(1)}$ in some order. What happens?

This is a situation where the good news is good and the bad news is better. If the spatial patterns are not too densely distributed in pattern space, in a sense that can be made precise (Grossberg, 1976a), then learning partitions the patterns into mutually exclusive and exhaustive subsets $P_1, P_2, \ldots, P_m$ such that every input pattern θ in P_j excites its recognition chunk $v_j^{(2)}$ with the maximal possible input, given the constraint that the LTM vector z_j is attracted to all the vectors f_θ of patterns θ in P_j.

Node $v_j^{(2)}$ is also activated by patterns θ that are weighted averages of patterns in P_j, even if these patterns θ are novel patterns that have never been experienced. Hence a generalization gradient exists across $\mathcal{F}^{(2)}$. The adaptive filter *projects* novel patterns into the classification set spanned by the patterns in P_j.

If a pattern θ is deformed so much that it crosses from one set P_j to another set P_k, then a rapid switch from choosing $v_j^{(2)}$ to choosing $v_k^{(2)}$ occurs. The boundaries between the sets P_j are categorical. Categorical perception can thus be anticipated whenever adaptive filtering interacts with sharply competitive tuning, not just in speech recognition experiments (Hary & Massaro, 1982; Pastore, 1981; Studdert-Kennedy, 1980).

The categorical boundaries are determined by how each input pattern is filtered by all the LTM vectors z_j, and by how all the dot product signals T_j fare in the global competition for STM activity. Consequently, practicing one pattern θ can recode the network's STM response to a novel pattern θ^* by changing the global balance between filtering and competition. This conclusion can be understood most easily by substituting Equations 38 and 39 into Equation 40. One then observes that the rate of change (d/dt) z_{ij} of each LTM trace depends on the global balance of all signals and all LTM traces, and thus on the entire history of the system as a whole.

Factors that promote adherence to or deviations from categorical perception are more subtle than these equations indicate. Section XXIII notes that an interplay of attentional factors with feature coding factors can cause the same network to react categorically or continuously to different experimental conditions. Such a result is not possible in the categorical perception model of Anderson et al. (1977), because the activities in that model must always reach a maximal or a minimal value (Section II).

XXIII. THE PROGRESSIVE SHARPENING OF MEMORY: TUNING PREWIRED PERCEPTUAL CATEGORIES

The requirement that $\mathcal{F}^{(2)}$ make an STM choice is clearly too strong. More generally, $\mathcal{F}^{(2)}$ possesses a tunable QT due to the fact that its competitive feedback signals are sigmoidal. Then only those signals T_j whose LTM vectors z_j are sufficiently parallel to an input pattern, within some range of tolerance determined by the QT, will cause suprathreshold STM reactions $x_j^{(2)}$. In this case, the competitive dynamics of $\mathcal{F}^{(2)}$ can be approximated by a rule of the form

$$x_j^{(2)} = \begin{cases} \dfrac{\mathrm{h}(T_j)}{\sum\limits_{T_k \geq \varepsilon} \mathrm{h}(T_k)} & \text{if} \quad T_j \geq \varepsilon \\ 0, & \text{if} \quad T_j < \varepsilon \end{cases} \tag{41}$$

instead of Equation 39. In Equation 41, the inequality $T_j \geq \varepsilon$ says that the dot product input to $v_j^{(2)}$ exceeds the QT of $\mathcal{F}^{(2)}$. The function $\mathrm{h}(T_j)$ approximates the contrast-enhancing action of sigmoid signaling within $\mathcal{F}^{(2)}$. The ratio of $\mathrm{h}(T_j)$ to $\sum_{T_k \geq \varepsilon} \mathrm{h}(T_k)$ approximates the normalization property of $\mathcal{F}^{(2)}$.

Due to Equation 41, a larger input T_j than T_k causes a larger STM reaction $x_j^{(2)}$ than $x_k^{(2)}$. By Equation 40, a larger value of $x_j^{(2)}$ than $x_k^{(2)}$ causes faster conditioning of the LTM vector z_j than of z_k. Faster conditioning of z_j causes $h(T_j)$ to be relatively larger than $h(T_k)$ on later learning trials than on earlier learning trials. Due to the normalization property, relatively more of the total activity of $\mathcal{F}^{(2)}$ will be concentrated at $x_j^{(2)}$ than was true on earlier learning trials. The relative advantage of $x_j^{(2)}$ is then translated into relatively faster conditioning of the LTM vector z_j. This feedback exchange between STM and LTM continues until the process equilibrates, if indeed it does. As a result of this exchange, the critical features within the filter $T = (T_1, T_2, \ldots, T_m)$ eventually overwhelm less salient features within the STM representation across $\mathcal{F}^{(2)}$ of input patterns to $\mathcal{F}^{(1)}$. Representations can thus be sharpened, or progressively tuned, due to a feedback exchange between "slow" adaptive coding and "fast" competition.

The tendency to sharpen representations due to training leads to learned codes with context-sensitive properties. This is because the critical features that code a given input pattern are determined by all of the input patterns that the network ever experiences. The ultimate representation of a single "word" in the network's input vocabulary thus depends on the entire "language" being learned, despite the fact that prewired connections in the signaling pathways from $\mathcal{F}^{(1)}$ to $\mathcal{F}^{(2)}$, and within $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$, constrain the features that will go into these representations.

Other sources of complexity are due to the fact that Equations 38, 40, and 41 approximate only the most elementary aspects of the learning process. The filter in Equation 38 often contains parameters P_{ij}, as in

$$T_j = \sum_{i=1}^{n} f(x_i^{(1)}) P_{ij} z_{ij}, \tag{42}$$

which determine prewired *positional gradients* from $\mathcal{F}^{(1)}$ to $\mathcal{F}^{(2)}$. These positional gradients break up the filtering of an input pattern into sets of partially overlapping channels. *Some* choice of prewired connections P_{ij} is needed to even define a filter such as that in Equation 38 or 42. Thus "tuning an adaptive filter" always means "tuning the prewired positional gradients of an adaptive filter." Infants may thus be "able to perceive a wide variety of phonetic contrasts long before they actually produce these contrasts in their own babbling" (Jusczyk, 1981, p. 156). The fact that developmental tuning may alter the LTM traces z_{ij} in Equation 42 in no way invalidates the ability of the prewired gradients P_{ij} in the equation to constrain the perceptual categories that tuning refines, both before and after tuning takes place.

The special choice $P_{ij} = 1$ for all i and j in Equation 38 describes the simplest (but an unrealistic) case in which all filters T_j receive equal prewired connections, but possibly different initial LTM traces, from all nodes v_i. In subsequent formulas, all P_{ij} will be set equal to 1 for notational simplicity. However, the need to choose nonuniform P_{ij} in general should not be forgotten.

Other simplifying assumptions must also be generalized in order to deal with realistic cases. The rule in Equation 41 for competition ignores many subtleties of how one competitive design can determine a different STM transformation than another. This rule also ignores the fact that the input pattern T to $\mathcal{F}^{(2)}$ is transformed into a pattern of STM traces across $\mathcal{F}^{(2)}$ before this STM pattern, not the input pattern itself, is further transformed by competitive feedback interactions within $\mathcal{F}^{(2)}$. Despite these shortcomings, the robust tendency for memory to sharpen progressively due to experience is clarified by these examples (Cermak & Craik, 1979; Squire, Cohen, & Nadel, 1982).

The degree of representational sharpening can be manipulated at will by varying the QT of $\mathcal{F}^{(2)}$. A high QT will cause sharply tuned codes to evolve from $\mathcal{F}^{(1)}$ to $\mathcal{F}^{(2)}$, as in Equation 39. A lower QT will enable a more diffusely distributed map to be learned from $\mathcal{F}^{(1)}$ to $\mathcal{F}^{(2)}$. Such a diffuse map may be protected from noise amplification by the use of sigmoidal competitive signaling at every processing stage that is capable of STM storage. If, however, the QT is chosen so small that fluctuations in internal cellular noise or in nonspecific arousal can exceed the QT, then the network STM and LTM can both be pathologically destabilized. A network in which two successive stages of filtering $\mathcal{F}^{(1)} \rightarrow \mathcal{F}^{(2)} \rightarrow \mathcal{F}^{(3)}$ occur—where the first stage $\mathcal{F}^{(1)} \rightarrow \mathcal{F}^{(2)}$ generates a diffuse map and the second stage $\mathcal{F}^{(2)} \rightarrow \mathcal{F}^{(3)}$ generates a sharply tuned map—is capable of computing significant global invariants of the input patterns to $\mathcal{F}^{(1)}$ (Fukushima, 1980; Grossberg, 1978e).

XXIV. STABILIZING THE CODING OF LARGE VOCABULARIES: TOP-DOWN EXPECTANCIES AND STM RESET BY UNEXPECTED EVENTS

Now for the bad news. If the number of input patterns to be coded is large relative to the number of nodes in $\mathcal{F}^{(2)}$, and if these input patterns are densely distributed in pattern space, then no temporally stable code may develop across $\mathcal{F}^{(2)}$ using only the interactions of the previous section (Grossberg, 1976a). In other words, the STM pattern across $\mathcal{F}^{(2)}$ that is caused by a fixed input pattern can persistently change through time due to the network's adaptive reaction to the other input patterns. Thc effort to overcome this catastrophe led to the introduction of the *adaptive resonance theory* (Grossberg, 1976b). I refer the reader to previous articles (Grossberg, 1982b, 1982d, 1984a) for a more thorough analysis of these results.

The main observation needed here is that a developing code can always be temporally stabilized by the action of conditionable top-down templates or feedback expectancies. This fact sheds new light on results which have suggested a role for feedback templates in a diverse body of data, including data about phonemic restoration, word superiority effects, visual pattern completion effects, and olfactory coding, to name a few (Dodwell, 1975; Foss & Blank, 1980; Freeman, 1979; Johnston & McClelland, 1974; Lanze, Weisstein, & Harris, 1982; Marslen-Wilson, 1975; Marslen-Wilson & Welsh, 1978; Rumelhart & McClelland, 1982; Warren, 1970; Warren & Obusek, 1971). My theory suggests that top-down templates are a universal computational design in all neural subsystems capable of achieving temporally stable adaptation in response to a complex input environment. The theory also identifies the mechanisms needed to achieve temporally stable adaptation. Becausc many articlcs that usc top-down mechanisms consider only performance issues rather than a composite of learning and performance issues, they do not provide a sufficient indication of why top-down expectancies exist or how they work.

My theory consider the basic issue of how a network can buffer the internal representations that it has already self-organized against the destabilizing effects of behaviorally irrelevant environmental fluctuations and yet adapt rapidly in response to novel environmental events which are crucial to its survival. To do this, the network must know the difference between and be able to differentially process both expected and unexpected events. I trace this ability to the properties of two complementary subsystems: an orienting subsystem and an attentional subsystem. Figures 6.15 and 6.16 summarize how these two types of subsystems operate.

In both figures, I assume that an active STM pattern is reverberating across certain nodes in the $(i + 1)$st field $\mathcal{F}^{(i+1)}$ of a coding hierarchy. These active nodes are emitting conditioned feedback signals to the previous stage $\mathcal{F}^{(i)}$ in this hierarchy. The total pattern E of these feedback signals represents the pattern that the active nodes in $\mathcal{F}^{(i+1)}$ collectively "expect" to find across $\mathcal{F}^{(i)}$ due to prior learning trials on which these nodes sampled the STM patterns across $\mathcal{F}^{(i)}$ via an associative process akin to outstar learning. More precisely, an expectancy E is a vector $E = (E_1, E_2, \ldots, E_n)$ such that $E_k = \Sigma_{j=1}^{m} S_j z_{jk}^{(i)}$ where S_j is the sampling

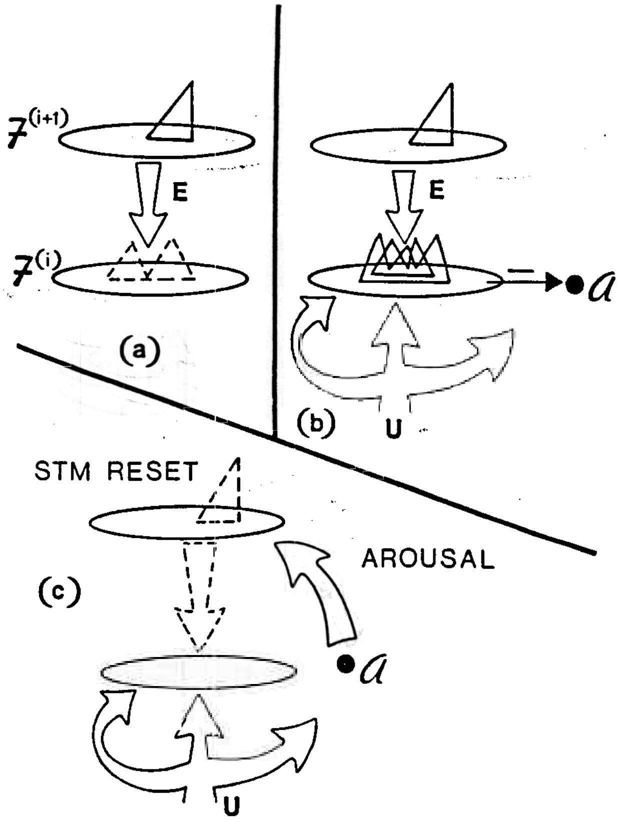

Figure 6.15 Reaction of attentional and orienting subsystems to an unexpected event: (a) A subliminal top-down expectancy E at $\mathcal{F}^{(1)}$ is maintained by a supraliminal STM pattern across $\mathcal{F}^{(2)}$. (b) The input pattern U nonspecifically sensitizes $\mathcal{F}^{(1)}$ as it instates itself across $\mathcal{F}^{(1)}$. The input also sends an activating signal to the nonspecific arousal source α. (c) The event U is unexpected because it mismatches E across $\mathcal{F}^{(1)}$. The mismatch inhibits STM activity across $\mathcal{F}^{(1)}$ and disinhibits α. This in turn releases a nonspecific arousal pulse that rapidly resets STM across $\mathcal{F}^{(2)}$ before adventitious recoding of the LTM can occur and drives a search of associative memory until a better match can be achieved.

signal emitted from $v_j^{(i+1)}$ and $z_{jk}^{(i)}$ is the LTM trace in the synaptic knobs of the pathway e_{jk} from $v_j^{(i+1)}$ to $v_k^{(i)}$.

I assume that the feedback signals E bias $\mathcal{F}^{(i)}$ by subliminally activating the STM traces across $\mathcal{F}^{(i)}$. Only a subliminal reaction is generated by the expectancy because the QT of $\mathcal{F}^{(i)}$ is assumed to be controlled by a nonspecific shunting signal, as in Equation 30. Although the expectancy E is active, the QT of $\mathcal{F}^{(i)}$ is too high for E to cause a supraliminal STM reaction.

A feedforward input pattern U from $\mathcal{F}^{(i-1)}$ to $\mathcal{F}^{(i)}$ has two effects on $\mathcal{F}^{(i)}$. It delivers the specific input pattern U and activates the nonspecific shunting signal that lowers the QT of $\mathcal{F}^{(i)}$. The conjoint action of U and E then determines the STM pattern elicited by U across $\mathcal{F}^{(i)}$.

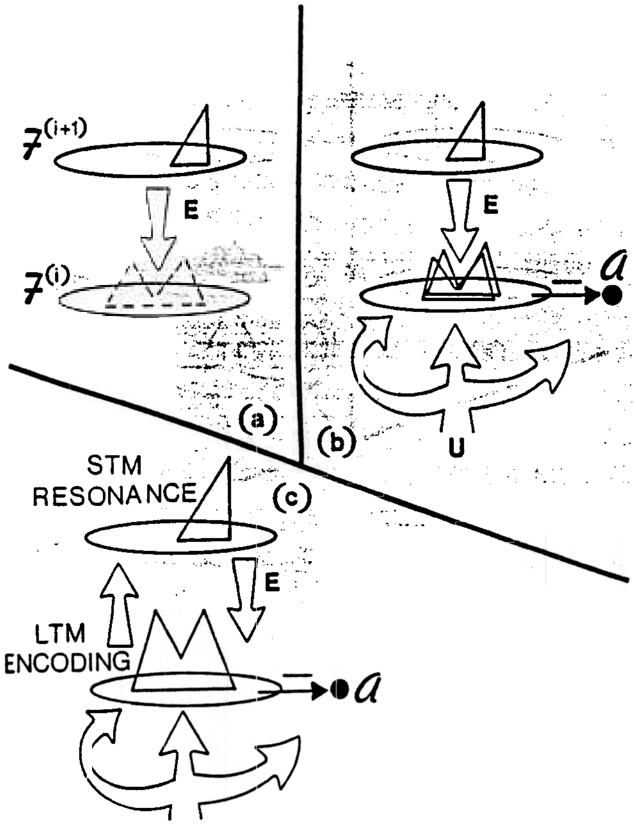

Figure 6.16 Reaction of attentional and orienting subsystems to an expected event: (a) A subliminal top-down expectancy E at $\mathcal{F}^{(1)}$ is maintained by a supraliminal STM pattern across $\mathcal{F}^{(2)}$. (b) The input pattern U nonspecifically sensitizes $\mathcal{F}^{(1)}$ as it instates itself across $\mathcal{F}^{(1)}$. The input also sends an activating signal to the nonspecific arousal source $\mathcal{A}$. (c) The event U is expected because it approximately matches E across $\mathcal{F}^{(1)}$; that is, it falls within the hysteresis boundary defined by E. This match amplifies patterned STM activity across $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ into a resonant STM code capable of being encoded in LTM.

It is worth noting at this point that any other input source capable of turning on the nonspecific shunting signal to $\mathcal{F}^{(i)}$ could lower its QT and thereby bootstrap the expectancy signals into a supraliminal STM pattern even in the absence of a feedforward input pattern. I believe that fantasy activities such as internal thinking and the recall of music can be maintained in this fashion.

We now consider how unexpected versus expected input patterns are differentially processed by this network. In Figure 6.15b, an unexpected input pattern U is delivered to $\mathcal{F}^{(i)}$ from $\mathcal{F}^{(i-1)}$. The pattern U is unexpected in the sense that the feedback template E and the unexpected input U are mismatched across $\mathcal{F}^{(i)}$. The concept of mismatch is a technical concept whose instantiation is a property of the interactions within $\mathcal{F}^{(i)}$ (Grossberg, 1980c, Appendix C; Carpenter and Grossberg, 1986a, 1986b).

For present purposes, we need to know only that a mismatch between two input patterns at $\mathcal{F}^{(i)}$ quickly attenuates STM activity across $\mathcal{F}^{(i)}$ (sec Figure 6.15c).

Just as the input pattern U activates a nonspecific gain control mechanism within $\mathcal{F}^{(i)}$, it also delivers an input to the orienting subsystem $\mathcal{A}$. Because each node in $\mathcal{F}^{(i)}$ sends an inhibitory pathway to $\mathcal{A}$(Figure 6.15b), suprathreshold STM activity anywhere across $\mathcal{F}^{(i)}$ can inhibit the input due to U at $\mathcal{A}$. If mismatch across $\mathcal{F}^{(i)}$ occurs, however, inhibitory signals from $\mathcal{F}^{(i)}$ to $\mathcal{A}$ are attenuated. Consequently, U's input to $\mathcal{A}$ is capable of unleashing a nonspecific signal to $\mathcal{F}^{(i+1)}$, which acts quickly to reset STM activity across $\mathcal{F}^{(i+1)}$.

One of the properties of STM reset is to selectively inhibit the active nodes across $\mathcal{F}^{(i+1)}$ that read out the incorrect expectancy of which event was about to happen. STM reset thus initiates a search for a more appropriate code with which to handle the unexpected situation. An equally important effect of inhibiting the active nodes in $\mathcal{F}^{(i+1)}$ is to prevent these nodes from being recoded by the incorrect pattern U at $\mathcal{F}^{(i)}$. STM reset shuts off the active STM traces $x_j^{(i+1)}$ across $\mathcal{F}^{(i+1)}$ so quickly that the slowly varying LTM traces $z_{kj}^{(i)}$ from $\mathcal{F}^{(i)}$ to $\mathcal{F}^{(i+1)}$ cannot be recoded via the LTM law

$$\frac{d}{dt} z_{kj}^{(i)} = [-z_{kj}^{(i)} + U_k]x_j^{(i+1)}.$$

The top-down expectancy thus buffers its activating chunks from adventitious recoding.

XXV. EXPECTANCY MATCHING AND ADAPTIVE RESONANCE

In Figure 6.16, the pattern U to $\mathcal{F}^{(i)}$ is expected. This means that the top-down expectancy E and the bottom-up pattern U match across $\mathcal{F}^{(i)}$. This notion of matching is also a technical concept that is instantiated by interactions within $\mathcal{F}^{(i)}$. When a pair of patterns match in this sense, the network can energetically amplify the matched pattern across $\mathcal{F}^{(i)}$ (see Figure 6.16b). These amplified activities cause amplified signals to pass from $\mathcal{F}^{(i)}$ to $\mathcal{F}^{(i+1)}$ (Figure 6.16c). The STM pattern across $\mathcal{F}^{(i+1)}$ is then amplified and thereupon amplifies the feedback signals from $\mathcal{F}^{(i+1)}$ to $\mathcal{F}^{(i)}$. This process of mutual amplification causes the STM patterns across $\mathcal{F}^{(i)}$ and $\mathcal{F}^{(i+1)}$ to be locked into a sustained STM resonance that represents a context-sensitive encoding of the expected pattern U. The resonant STM patterns can be encoded by the LTM traces in the pathways between $\mathcal{F}^{(i)}$

and $\mathcal{F}^{(i+1)}$ because these STM patterns are not rapidly inhibited by an STM reset event. Because STM resonance leads to LTM encoding, I call this dynamical event an *adaptive resonance.*

XXVI. THE PROCESSING OF NOVEL EVENTS: PATTERN COMPLETION VERSUS SEARCH OF ASSOCIATIVE MEMORY

A novel input pattern U can elicit two different types of network reaction, depending on whether U triggers STM resonance or STM reset. When a novel event U is filtered via feedforward $\mathcal{F}^{(i)} \rightarrow \mathcal{F}^{(i+1)}$ signaling, it may activate a feedback expectancy E via feedback $\mathcal{F}^{(i+1)} \rightarrow \mathcal{F}^{(i)}$ signaling which, although not the same pattern as U, is sufficiently like U to generate an approximate match across $\mathcal{F}^{(i)}$. This can happen because the QT of $\mathcal{F}^{(i)}$ determines a flexible criterion of how similar two patterns must be to prevent the inhibition of all STM activity across $\mathcal{F}^{(i)}$. A large QT implies a strict criterion, whereas a low QT implies a weak criterion. If two patterns are matched well enough for some populations in $\mathcal{F}^{(i)}$ to exceed the QT, then STM resonance will occur and the orienting reaction will be inhibited.

Because U is filtered by feedforward signaling, and because E reads out the optimal pattern that the active chunks across $\mathcal{F}^{(i+1)}$ have previously learned, E will deform the STM reaction across $\mathcal{F}^{(i)}$ that would have occurred to U alone toward a resonant encoding that completes U using the optimal data E. This consequence of buffering LTM codes against adventitious recoding is, I believe, a major source of Gestaltlike pattern completion effects, such as phonemic restoration, word superiority effects, and the like. Grossberg and Stone (1986a) develop such concepts to analyze data about word recognition and recall.

By contrast, if U is so different from E that the QT causes STM suppression across $\mathcal{F}^{(i)}$, then the orienting subsystem will be activated and a rapid parallel search of associative memory will ensue. To understand how such a search is regulated, one needs to analyze how a nonspecific arousal pulse to $\mathcal{F}^{(i+1)}$ can selectively inhibit only the active nodes across $\mathcal{F}^{(i+1)}$ and spare inactive nodes for subsequent encoding of the unexpected event.

This property is instantiated by expanding the design of $\mathcal{F}^{(i+1)}$, as well as all other network levels that are capable of STM reset, in the following fashion. All nodes that have heretofore been posited in the competitive networks are *on-cells;* they are turned on by inputs. Now I supplement the on-cell competitive networks with apposing off-cell competitive networks such that offset of an input to an on-cell triggers a transient activation of its corresponding off-cell. Such an activation is called an *antagonistic rebound.*

Antagonistic rebound at an off-cell in response to offset of an on-cell input can be achieved due to three mechanisms acting together: (1) All the inputs to both the on-cell channel and the off-cell channel are gated by slowly accumulating transmitter substances that generate output signals by being released at a rate proportional to input strength times the amount of available transmitter. (2) The inputs to on-cells and off-cells are of two types: specific inputs that selectively activate an on-cell channel or an off-cell channel, but not both, and nonspecific inputs that activate both on-cell and off-cell channels equally. (3) The gated signals in both the on-channel and off-channel compete before the net gated inputs activate the on-cell or the off-cell, but not both. The network module that carries out these computations is called a *gated dipole* (Figure 6.17). One proves that if a sufficiently large increment in nonspecific arousal occurs while an on-cell is active, this increment can cause an antagonistic rebound that rapidly shuts off the on-cell's STM trace by exciting the corresponding off-cell. This rebound is part of the STM reset event.

The antagonistic rebounds in gated dipoles are due to the fact that unequal inputs to the on-cell and off-cell cause their respective transmitter gates to be depleted, or *habituated,* at unequal rates. When a rapid change in input patterning occurs, it is gated by the more slowly varying transmitter levels. A mathematical analysis shows that either a rapid offset of the specific on-cell input or a rapid increase of nonspecific arousal can cause an antagonistic rebound due to the imbalance in transmitter habituation across the on-cell and off-cell channels (Grossberg, 1972b, 1975, 1980c, 1981b, 1984a).

Once a subfield of on-cells is inhibited by dipole rebounds, they remain inhibited for a while due to the slow recovery rate of the transmitter gates. Only a subset of nodes in $\mathcal{F}^{(i+1)}$ can therefore respond to the filtered signals from $\mathcal{F}^{(i)}$ in the next time interval. If another mismatch occurs, more nodes in $\mathcal{F}^{(i+1)}$ are inhibited. As the search continues, the normalized STM patterns across $\mathcal{F}^{(i+1)}$ contract rapidly onto a final subset of $\mathcal{F}^{(i+1)}$ nodes. The STM pattern across this final subset of nodes is used to condition the filter of its corresponding pathways or to stabilize the already conditioned pathways via an adaptive resonance.

One of the intriguing facts about searching associative memory in this way is that transmitter habituation is one of the important mechanisms. Habituation acts in my theory to regulate active memory buffering and adaptive encoding processes; it is not just a passive result of "use," "fatigue," or other classical notions.

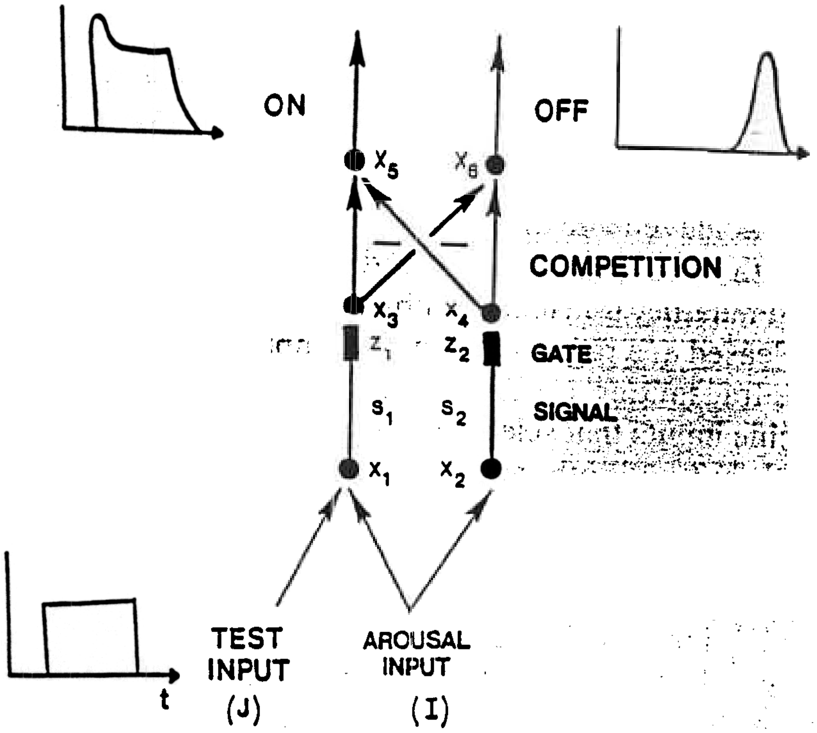

Figure 6.17 Reaction of a feedforward gated dipole to phasic onset and offset of a cue: The phasic test input J and the arousal input I add in the on-channel, thereby activating the STM trace x_1. The arousal input I activates the STM trace x_2 in the off-channel. Since $I + J > I$, $x_1 > x_2$. Traces x_1 and x_2 elicit signals $f(x_1)$ and $f(x_2)$. Because $x_1 > x_2$, $f(x_1) > f(x_2)$. Each signal is gated by a transmitter z_1 or z_2 (in the square synapses). Transmitters z_1 and z_2 are released at rates proportional to $f(x_1)z_1$ and $f(x_2)z_2$, respectively. The gated signals $f(x_1)z_1$ and $f(x_2)z_2$ in turn activate x_3 and x_4, respectively, and $x_3 > x_4$. Both x_3 and x_4 excite their own channel and inhibit the other channel. After competition acts, an output from the on-channel is elicited that habituates through time. A rapid offset of J causes $x_1 = x_2$ and $f(x_1) = f(x_2)$. However, $z_1 < z_2$ due to greater prior depletion, or habituation, of the on-channel by J, and to the slow reaction rate of z_1 and z_2 relative to x_1 and x_2. Thus $f(x_1)z_1 < f(x_2)z_2$ and $x_3 < x_4$. After competition acts, the off-channel wins. Gradually, $z_1 = z_2$ in response to the equal values of x_1 and x_2. Then $f(x_1)z_1 = f(x_2)z_2$, so the competition shuts off all dipole output. The same mechanism causes on off-rebound in response to a rapid increment in I while J is active. In a feedback gated dipole, output from x_5 reexcites x_1, and output from x_6 reexcites x_2. In a dipole field, the positive feedback loops $x_1 \leftrightarrow x_5$ and $x_2 \leftrightarrow x_6$ form the on-centers of competitive shunting networks that join on-cells to on-cells and off-cells to off-cells. Such networks exhibit STM properties such an contrast enhancement and normalization modulated by the slow habituation of activated transmitter gates.

XXVII. RECOGNITION, AUTOMATICITY, PRIMES, AND CAPACITY

These processes are reflected in many types of data. Analyses and predictions about some of these data are found in Grossberg (1980c, 1982b, 1982c). Carpenter and Grossberg (1986a, 1986b) describe extensive computer simulations of how adaptive resonance theory mechanisms can self-organize a self-stabilizing recognition code in response to an arbitrary list of input patterns. The following remarks summarize some of the other types of data that may be clarified by such processes.

Recent studies of recognition memory point out: "Complex elaborate encoding . . . can be utilized to enhance recognition only when the test conditions permit a reinstatement of the original encoding context" (Fisher & Craik, 1980, p. 400). In a similar vein, other studies support the idea that "the direction of priming effects . . . depend[s] upon the validity of the prime as a predictor of the probe stimulus" (Myers & Lorch, 1980, p. 405). This type of effect holds at every level of reciprocal signaling in the encoding hierarchy, because a particular pattern of feedforward chunking is wed to a characteristic pattern of template feedback. An active pattern of template feedback leads to rapid resonant matching only when it meets with compatible feedforward input patterns. The resonant process is interpreted to be the attentive moment, or recognition event, that groups individual nodal activities into a unitary percept. The "priming" feedback template leads to enhanced recognition only if the "probe" input pattern is sufficiently similar to trigger a resonant match.

This view of template matching and associative memory search suggests a different way to think about automatic versus capacity-limited processing than that of researchers like Norman and Bobrow (1975) or Posner and Snyder (1975). Capacity limitations alone do not determine whether very fast inhibition will slow down reaction times in a search situation. Mismatch can trigger rapid STM reset and associative search, leading to an increase in reaction time, under the same capacity limitations that would speed reaction time if a match were to occur. An increased reaction time is not due merely to a capacity limitation that excites two mutually inhibitory nodes, because mutual inhibition can subserve either a match or a mismatch.

At bottom, the traditional discussion of increased reaction time due to capacity-limited inhibitory effects follows from an inadequate choice of the functional unit of network processing. The functional unit is not the activation of a single node, nor a "spreading activation" among individual nodes. Rather, it is a coherent pattern across a field of nodes. This viewpoint is compatible with the results of Myers and Lorch (1980, p. 405), who showed, among other things: "Reaction time to decide that a sentence was true or false was longer if the preceding prime was a word that was unrelated to the probe than if the prime was the word 'blank' " at prime-to-probe intervals as short as 250 ms.

How can a more direct test of behaviorally unobservable reset and search routines be made? One possible way may be to adapt techniques for measuring the N200 and P300 evoked potentials to letter, word, and sentence recognition and verification tasks. The theory suggests that every mismatch event will elicit the mismatch negativity component of the N200 at $\mathscr{A}$ and that every subsequent nonspecific arousal burst will elicit a P300 at $\mathcal{F}^{(2)}$ (Grossberg, 1984a; Karrer, Cohen, & Tueting, 1984).

XXVIII. ANCHORS, AUDITORY CONTRAST, AND SELECTIVE ADAPTATION

Numerous studies of contextual effects in vowel perception have attempted to distinguish between "feature detector fatigue" and "auditory contrast" explanations of how a categorical boundary can shift during a selective adaptation or anchoring experiment. Sawusch and Nusbaum (1979) have interpreted their results as favoring an auditory contrast explanation, because even widely spaced repetitions of an anchor vowel elicit a significant shift in the boundary. The mechanisms of template feedback, STM resonance, STM reset, and transmitter habituation may shed some further light on this discussion by suggesting some new experimental tests of these ideas.

Sawusch and Nusbaum (1979, p. 301) discuss auditory contrast in terms of "incorporating the influence of both information from immediately preceding stimuli (auditory memory) and prototypes in long-term memory into one unified auditory ground against which new stimuli are compared." I represent the "auditory memory" by a pattern reverberating in STM across a field $\mathcal{F}^{(i+1)}$ of nodes, the "prototypes in long-term memory" by LTM traces in the active feedback template pathways from $\mathcal{F}^{(i+1)}$ to $\mathcal{F}^{(i)}$, and the "unified auditory ground" by a subliminal feedback expectancy E across $\mathcal{F}^{(i)}$. This interpretation immediately raises several questions.

Why does an input U that mismatches E cause a boundary shift *away* from the event represented by E? In the framework presented earlier, the answer is: The mismatch event actively inhibits the active nodes across $\mathcal{F}^{(i+1)}$ which represent E. In the time interval after this STM reset event, the pattern U will be encoded by a renormalized field $\mathcal{F}^{(i+1)}$ in which the nodes that encoded E remain inhibited. A similar combination of mismatch-then-reset has been used to discuss bistable visual illusions such as those that occur during the viewing of Necker's cube (Grossberg, 1980c). I suggest that a P300-evoked potential will occur at the moment of switching.

As in the discussion of matching probes to primes, a stronger anchoring effect may cause a faster reaction time when U equals E and a slower reaction time when U mismatches E. Another type of test would start with an event U that equals E and would gradually cause U to mismatch E using a temporally dense series of successive presentations of slightly changing Us. The hysteresis boundary that causes perseveration of the anchor percept may get broader as the anchoring effect is made stronger, even though a stronger anchoring effect causes a larger shift when a discrete event U mismatches E. A P300 may also occur when the hysteresis boundary is exceeded by U. Thereafter, the percept may again be shifted away from E.

The latter test may be confounded by the fact that a dense series of Us may cause persistent STM reverberation of the anchor representation across $\mathcal{F}^{(i+1)}$. Such a reverberation may progressively habituate the transmitter gates in the reverberating pathways. In this way, "fatigue" may enter even an auditory contrast explanation, albeit fatigue of a nonclassical kind. If significant habituation does occur, then the shift due to STM reset may become smaller as a function of longer storage in auditory memory, but this effect would be compensated for by a direct renormalization of the STM response of $\mathcal{F}^{(i+1)}$ to U as a result of habituation. Such habituation effects may occur on a surprisingly long time scale, because a slow transmitter accumulation rate is needed to regulate the search of associative memory.

Sawusch and Nusbaum (1979) suggest that adaptation-level theory (Helson, 1964; Restle, 1978) may be used as a mathematical framework to explain selective adaptation and anchoring effects. I believe that this is a correct intuition. Shunting competitive networks possess an adaptation level that is the basis for their matching properties (Grossberg, 1980c, Appendix C; 1983, Section 22). A feedback template E has the effect of biasing the adaptation level and thereby producing a different reaction to a feedforward input pattern U than would otherwise occur. However, this property of shunting networks controls only one step in the network's total reaction to a shifted input.

In this regard, Sawusch, Nusbaum, and Schwab (1980) show that anchoring by the vowel [i] (as in *beet*) or by [I] (as in *bit*) produce contrast effects in tests involving [i]–[I] vowel series by affecting different mechanisms. "Contrast effects for an [i] anchor were found to be largely the result of changes in sensitivity between various vowel pairs . . . the [i] anchoring effect occurs prior to phonetic labeling. This is clearly the case, since [i] anchoring was found to increase discriminability within the [i] category . . . The [i] anchor seems to alter or retune the prototype space." By contrast, "the [I]-anchor effects were largely the result of criterion shifts . . . The auditory ground would reflect two sources of information: prototype information from long-term memory and certain information from the stimulus being presented . . . some form of auditory memory which contains information about the quality of the stimulus may underlie the changes in criterion for [I] anchoring" (Sawusch et al., 1980, p. 431).

Within the present theory, the changes in discriminability whereby the [i] anchor "retunes the prototype space" may be interpreted as an [i]-induced shift in some of the LTM vectors that form part of the auditory adaptive filter (Sections XXII–XXIII). This LTM tuning process occurs prior to the stage of "phonetic labeling," or STM competition in auditory memory, and changes the outcome of the phonetic competition by altering the pattern of filtered inputs on which the competition feeds.

Both the [i] vowel and the [I] vowel can bias the adaptation level by reading their subliminal feedback templates out of auditory memory. This type of top-down bias can create criterion shifts without redistributing the bottom-up LTM vectors in the adaptive filter that control relative sensitivity. If no change in the adaptive filter takes place, interference with auditory memory will reduce contrast effects due to anchoring. However, if an adaptive filter shift has occurred prior to interference with auditory memory, a large anchoring effect can still occur due to the direct effect on each trial of the shifted bottom-up filtering on top-down template readout. Sawusch et al. (1980) demonstrate an analogous effect. They partially interfere with auditory memory by embedding [i] and [I] in CVC syllables, such as [sis] and [sIs]. In this case, the anchoring effect of [sis] was significantly greater than that of [sIs].

XXIX. TRAINING OF ATTENTIONAL SET AND PERCEPTUAL CATEGORIES

Studdert-Kennedy (1980) has reviewed data that are compatible with this interpretation of the auditory ground. Spanish–English bilinguals can shift their boundaries by a change in language set within a single test (Elman, Diehl, & Buchwald, 1977). This shift can be formally accomplished in a network by activating nodes across $\mathcal{F}^{(i+1)}$ that read out different feedback templates.

Training enables subjects to shift categorical boundaries at will, thereby suggesting that "utilization of acoustic differences between speech stimuli may be determined primarily by attentional factors" (Carney, Widen, & Viemeister, 1977, p. 969). Such training may tune both the adaptive filters and the feedback templates of the subjects, much as American English speakers perceive an [r] to [l] continuum categorically, whereas Japanese speakers do not (Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975).

XXX. CIRCULAR REACTIONS, BABBLING, AND THE DEVELOPMENT OF AUDITORY–ARTICULATORY SPACE

Using the operations that have been sketched above, one can quantitatively discuss how neural networks initiate the process whereby their sensory and motor potentialities are integrated into a unitary system. The main concepts are that endogenously activated motor commands generate patterns of sensory feedback; that internal representations of the activated sensory and motor patterns are synthesized (learned, chunked) by adaptive tuning of coarsely prewired filters (feature detectors, positional gradients); and that the sensory internal representations are joined to their motor counterparts via a learned associative map. The sensory internal representations can also be tuned by sensory inputs other than sensory feedback as soon as external sensory inputs can command the attention-for example, lower the QT (Section XVIII)—of the sensory modality.

These concepts were first used in real-time network models to discuss how motivated and attentive behavior can be self-organized in a freely moving animal (Grossberg, 1971, 1972a, 1972b, 1975). Later, they were used to show how cognitive, attentive, and motivational mechanisms can interact to generate a consistent, goal-oriented sensory-motor plan (Grossberg, 1978e; reprinted in Grossberg, 1982d).

An earlier version of this approach is embodied in Piaget's concept of a *circular reaction* (Piaget, 1963). Then Fry (1966) emphasized the importance of the infant's babbling stage for the later development of normal speech (Marvilya, 1972), notably for the tuning of prewired adaptive sensory filters. The work of Marler and his colleagues (Marler, 1970; Marler & Peters, 1981) on the development of birdsong in sparrows has recognized the relevance of self-generated auditory feedback to the development of normal adult song. Similarly, the motor theory of speech perception recognized the intimate relationship between acoustic encoding and articulatory requirements (Cooper, 1979; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Studdert-Kennedy, 1978; Mann & Repp, 1981; Repp & Mann, 1981; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). Studdert-Kennedy (1975, 1980) has written about this approach with particular eloquence: "Only by carefully tracking the infant through its first two years of life shall we come to understand adult speech perception and, in particular, how speaking and listening establish their links at the base of the language system" (Studdert-Kennedy, 1980, p. 45). "The system follows the moment-to-moment acoustic flow, apprehending an auditory 'motion picture,' as it were, of the articulation" (p. 55).

This and the next section sketch a framework for analyzing how individual sensory and motor patterns are integrated. After that, I consider the deeper question of how temporal sequences of patterns are processed. To carry out this discussion, I use the notation $\mathcal{F}_M^{(i)}$ for the ith motor field in a coding hierarchy, and $\mathcal{F}_S^{(i)}$ for the ith sensory field in a coding hierarchy. The example of babbling in an infant can be used to intuitively fix ideas.

Suppose that an activity pattern across $\mathcal{F}_M^{(1)}$ represents a terminal motor map (TMM) of a motor act. Such a map specifies the terminal lengths of target muscles. It is often organized in a way that reflects the agonist–antagonistic organization of these muscles. Suppose that during a specified time interval, a series of TMMs are endogenously activated across $\mathcal{F}_M^{(1)}$, analogous to the babbling of simple sounds. The execution of such a TMM elicits sensory feedback, analogous to a sound, which is registered as an input pattern across $\mathcal{F}_S^{(1)}$ (Figure 6.18a).

As these unconditional events are taking place, they are accompanied by the following adaptive reactions. The active TMM is chunked by adaptive filtering from $\mathcal{F}_M^{(1)}$ to $\mathcal{F}_M^{(2)}$. This motor code in turn learns its corresponding TMM via the conditioning of its feedback template from $\mathcal{F}_M^{(2)}$ to $\mathcal{F}_M^{(1)}$. As this learning is taking place, the corresponding sensory feedback pattern across $\mathcal{F}_S^{(1)}$ is chunked by adaptive filtering from $\mathcal{F}_S^{(1)}$ to $\mathcal{F}_S^{(2)}$. Its sensory code learns the corresponding pattern of sensory features across $\mathcal{F}_S^{(1)}$ by the conditioning of its template feedback from $\mathcal{F}_S^{(2)}$ to $\mathcal{F}_S^{(1)}$ (Figure 6.18b).

Due to the simultaneous activity of the sensory and motor representations in $\mathcal{F}_S^{(2)}$ and $\mathcal{F}_M^{(2)}$, a map from $\mathcal{F}_S^{(2)}$ to $\mathcal{F}_M^{(2)}$ can be self-organized by associative pattern learning (Figure 6.18c). One of the quantitative issues of the theory concerns how diffuse or sharply tuned this map should be (Grossberg, 1978e, Sections 55–58).

The above network matches auditory to articulatory requirements in several ways. It preferentially tunes the sensory "feature detectors" that are activated most often by spoken sounds (Section XXII). It also maps the tuned internal representations of these sounds onto the motor commands that are capable of eliciting the sounds. The construction accomplishes this by associatively sampling the motor commands in the form that succeeded in executing the sounds through endogenous activation. Although the internal representations of sounds and motor commands may differ in many significant ways, they can be joined together by the common coin of pattern learning in associative networks. These patterns are the still pictures in the "motion picture" described by Studdert-Kennedy (1980).

The flow of activity in such a network is circular. It proceeds from $\mathcal{F}_M^{(1)}$

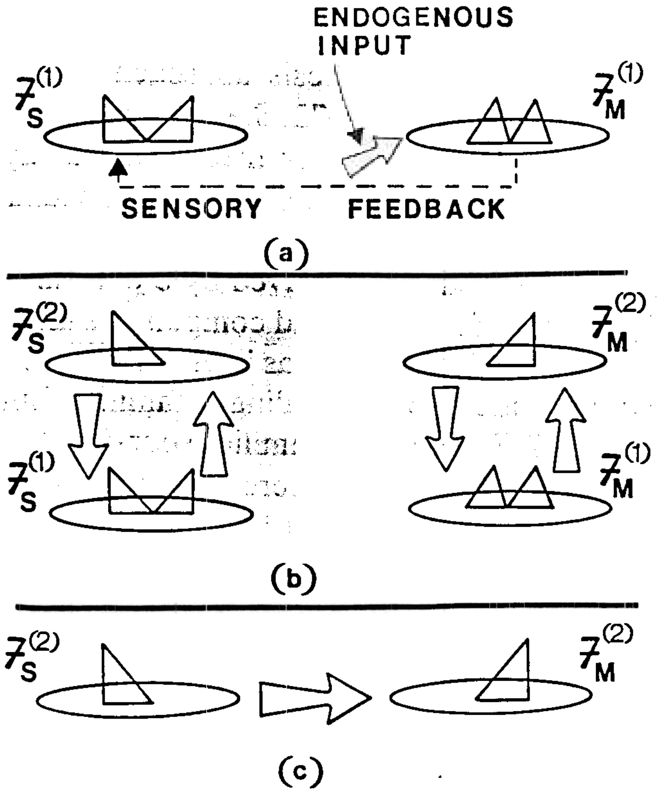

Figure 6.18 A circular reaction that matches acoustic encoding to articulatory requirements: (a) Endogenous motor commands across $\mathcal{F}_M^{(1)}$ elicit babbled sounds that are received as auditory feedback patterns across $\mathcal{F}_S^{(1)}$. (b) The motor commands and the auditory feedback patterns are chunked at $\mathcal{F}_M^{(2)}$ and $\mathcal{F}_S^{(2)}$ via bottom-up adaptive filtering in pathways $\mathcal{F}_M^{(1)} \rightarrow \mathcal{F}_M^{(2)}$ and $\mathcal{F}_S^{(1)} \rightarrow \mathcal{F}_S^{(2)}$, respectively. These chunks learn their generative motor commands and auditory patterns via top-down expectancy learning in pathways $\mathcal{F}_M^{(2)} \rightarrow \mathcal{F}_M^{(1)}$ and $\mathcal{F}_S^{(2)} \rightarrow \mathcal{F}_S^{(1)}$, respectively. (c) The sensory and motor chunks are joined together by an associative map $\mathcal{F}_S^{(2)} \rightarrow \mathcal{F}_M^{(2)}$. The learned map $\mathcal{F}_S^{(1)} \rightarrow \mathcal{F}_S^{(2)} \rightarrow \mathcal{F}_M^{(2)} \rightarrow \mathcal{F}_M^{(1)}$ completes the circular reaction and enables novel sounds to be imitated and then chunked by the same mechanisms.

to $\mathcal{F}_S{}^{(1)}$ via external sensory feedback, and in the reverse direction by a combination of adaptive filtering, associative mapping, and conditionable template feedback. This is a circular reaction, network-style. The completion of the circle by the internal network flow enables simple imitation to be accomplished via the learned map

$$\mathcal{F}_S{}^{(1)} \rightarrow \mathcal{F}_S{}^{(2)} \rightarrow \mathcal{F}_M{}^{(2)} \rightarrow \mathcal{F}_M{}^{(1)}.$$

XXXI. ANALYSIS-BY-SYNTHESIS AND THE IMITATION OF NOVEL EVENTS

After babbling stops, how does language continue to develop? In particular, how does a network learn to recognize and recall novel sounds other than those learned during the babbling phase? Part of this capability is built into the map in Equation 44 in a way that sheds light on the many successes of the analysis-by-synthesis approach to speech recognition (Halle & Stevens, 1962; Stevens, 1972; Stevens & Halle, 1964). The structure of the map suggests that motor theory and analysis-by-synthesis theory have probed different aspects of the same underlying physical process.

Suppose that a novel sound is received by $\mathcal{F}_S^{(1)}$. This sound is decomposed, or analyzed, into familiar sound components in the following way. The adaptive filter from $\mathcal{F}_S^{(1)}$ to $\mathcal{F}_S^{(2)}$ has been tuned by experience in such a way that the dot products corresponding to familiar sound patterns elicit larger inputs across $\mathcal{F}_S^{(2)}$ than do unfamiliar sound patterns, as in Section XXIII. This conclusion must be tempered by the fact that filter tuning of LTM traces z_{ij} is driven by the initial filtering of sound patterns by prewired positional gradients P_{ij}, as in Equation 42. The tuned adaptive filter analyzes the novel sound in a weighted combination of familiar sounds wherein the weights correspond to the relative activations of representational nodes across $\mathcal{F}_S^{(2)}$. Suppose that the associative map from $\mathcal{F}_S^{(2)}$ to $\mathcal{F}_M^{(2)}$ is diffuse. In this case the spatial pattern of weights that represent the novel sound is relayed as signal strengths from $\mathcal{F}_S^{(2)}$ to $\mathcal{F}_M^{(2)}$. Each familiar motor command at $\mathcal{F}_M^{(2)}$ is activated with an intensity corresponding to the size of the signal that it receives. All of the excited motor commands read out their TMMs to $\mathcal{F}_M^{(1)}$ with relative intensities corresponding to the relayed weights. The total read-out of familiar motor commands is then synthesized into a novel TMM across $\mathcal{F}_M^{(1)}$. The net effect of this analysis-by-synthesis process is to construct a novel TMM across $\mathcal{F}_M^{(1)}$ in response to a novel sound at $\mathcal{F}_S^{(1)}$.

The novel TMM needs to process the following *continuous mapping property:* It should elicit a sound that is more similar to the novel sound than to any of the familiar sounds in the network's repertoire. To achieve this property, continuous changes in the TMMs across $\mathcal{F}_M^{(1)}$ need to correspond to continuous changes in the auditory feedback patterns across $\mathcal{F}_S^{(1)}$. The continuous mapping property is most easily achieved by organizing auditory representations and motor representations in a topographic fashion.

If a novel TMM possessing the continuous mapping property is activated at $\mathcal{F}_M^{(1)}$ while the novel sound is still represented at $\mathcal{F}_S^{(1)}$, the network can build internal representations for both the sound and the TMM, as well as an associative map between these representations, just as it did for babbled sounds. As more internal representations and maps are built up, the network's ability to imitate and initiate novel sounds will become progressively refined. The metrical distances between a novel sound pattern or TMM pattern and the familiar sound or TMM patterns into which it is decomposed can be used as an intrinsic measure of the phenomenal complexity, or nonautomaticity, of a new behavior relative to a network code of familiar behaviors.

XXXII. A MOVING PICTURE OF CONTINUOUSLY INTERPOLATED TERMINAL MOTOR MAPS: COARTICULATION AND ARTICULATORY UNDERSHOOT

A topographic structuring of motor representations can be achieved by organizing these representations into agonist–antagonist pairs. The relative activations of such pairs can be signaled independently as part of a larger spatial pattern, much as individual articulators in the vocal tract, such as the lips, tongue, and vocal chords, can move with a large degree of independence (Darwin, 1976). A temporal sequence of TMMs from such a motor field is expressed as a sequence of spatial pattern outputs (Section V). Each spatial pattern controls a motor synergy of articulatory motions to intended positional targets (Section VI). The intrinsic organization of the articulatory system continuously interpolates the motion of the articulators between these targets.

The timing of component subpatterns within a sequence of spatial pattern TMMs can cause coarticulation to occur (Fowler, 1977). A high rate of emitting TMMs, due, say, to an increase in the gain of the read-out system by a nonspecific arousal increment (Section VII), can cause the next TMM to be instated before the last target has been attained. Rapid speech can thus be associated with articulatory undershoot and a corresponding acoustic undershoot (Lindblom, 1963; Miller, 1981).

XXXIII. A CONTEXT-SENSITIVE STM CODE FOR EVENT SEQUENCES

I now sketch some results about how sequences of events can be performed out of STM after a single presentation, and of how sequences of events can generate context-sensitive representations in LTM that are capable of accurately controlling planned, or predictive, behavior. These properties can be achieved by parallel mechanisms. No serial buffer is necessary.

Several classes of phenomena have been analyzed using these concepts, notably phenomena concerning free recall, letter and word recognition, and skilled motor behavior (Grossberg, 1978a, 1978e). A number of other authors have also discussed these phenomena using a network approach (e.g., MacKay, 1982; McClelland & Rumelhart, 1981; Norman, 1982; Rumelhart & McClelland, 1982; Rumelhart & Norman, 1982). Although I am in complete sympathy with these contributions, I believe that they have overlooked available principles and mechanisms that are essential for achieving better understanding of their targeted data. In the next few sections, I focus my discussion on how the functional unit of speech perception is self-organized by "an active continuous process" (Studdert-Kennedy, 1980), notably how backward effects, time-intensity tradeoff effects, and temporal integration processes can alter a speech percept (Miller & Liberman, 1979; Repp, 1979; Repp, Liberman, Eccardt, & Pesetsky, 1978; Schwab, Sawusch, & Nusbaum, 1981) in a manner that is difficult to explain using a computer model of speech perception (Levinson & Liberman, 1981). These ideas are supplemented by some mechanisms helpful in the analysis of rhythmic substrates of speech, skilled motor control, and musical performance (Fowler, 1977; Rumelhart & Norman, 1982; Shaffer, 1982; Studdert-Kennedy, 1980).

XXXIV. STABLE UNITIZATION AND TEMPORAL ORDER INFORMATION IN STM: THE LTM INVARIANCE PRINCIPLE

For simplicity, I begin by supposing that unitized item representations $v_1^{(3)}, v_2^{(3)}, \ldots, v_n^{(3)}$ in a field $\mathcal{F}^{(3)}$ are sequentially activated by a list of events $r_1, r_2, \ldots, r_n$. To fix ideas, the reader may suppose that the unitized representations are generated by adaptive filtering from either $\mathcal{F}_S^{(2)}$ or $\mathcal{F}_M^{(2)}$, since similar temporal order mechanisms are used in both sensory and motor modalities (Kimura, 1976; Kinsbourne & Hicks, 1978; Semmes, 1968; Studdert-Kennedy, 1980).

Suppose that a certain number of nodes $v_1^{(3)}, v_2^{(3)}, \ldots, v_i^{(3)}$ have been activated by the sublist $r_1, r_2, \ldots, r_i$ and therefore have active STM traces at a given time t_i. At this moment, the set of active STM traces defines a spatial pattern. Had the same sublist been presented in a different order, a different STM pattern would exist across the same set of nodes. Thus the active STM pattern encodes temporal order information across the item representations.

To achieve a correct read-out of temporal order information directly from STM, a primacy effect in STM,

$$x_1^{(3)} > x_2^{(3)} > \cdots > x_i^{(3)}, \tag{45}$$

is desired, as in Equation 25. Section XII shows how a temporal series of recency effects in STM can elicit a learned read-out from LTM of a primacy effect in STM. We now consider how a primacy effect in STM can sometimes be caused *directly* by experimental inputs, yet also how a recency effect in STM can sometimes be caused by experimental inputs, thereby leading to order errors in the read-out of items from STM.

To understand this issue, I abandon all homunculi and consider how the evolving STM pattern can be encoded in LTM in a temporally stable fashion by the adaptive filter from $\mathcal{F}^{(3)}$ to $\mathcal{F}^{(4)}$. This adaptive filter groups together, or unitizes, sublists of the items that are simultaneously stored in STM at $\mathcal{F}^{(3)}$. The STM pattern at $\mathcal{F}^{(4)}$ codes as unitized sublist chunks those item groupings that are salient to the network when a prescribed list of items is stored at $\mathcal{F}^{(3)}$ (Figure 6.19). Thus I now consider laws for storing individual items in STM at $\mathcal{F}^{(3)}$ which enable the LTM unitization process to proceed in a stable fashion within the adaptive filter from $\mathcal{F}^{(3)}$ to $\mathcal{F}^{(4)}$. In short, I constrain STM to be compatible with LTM. This is a

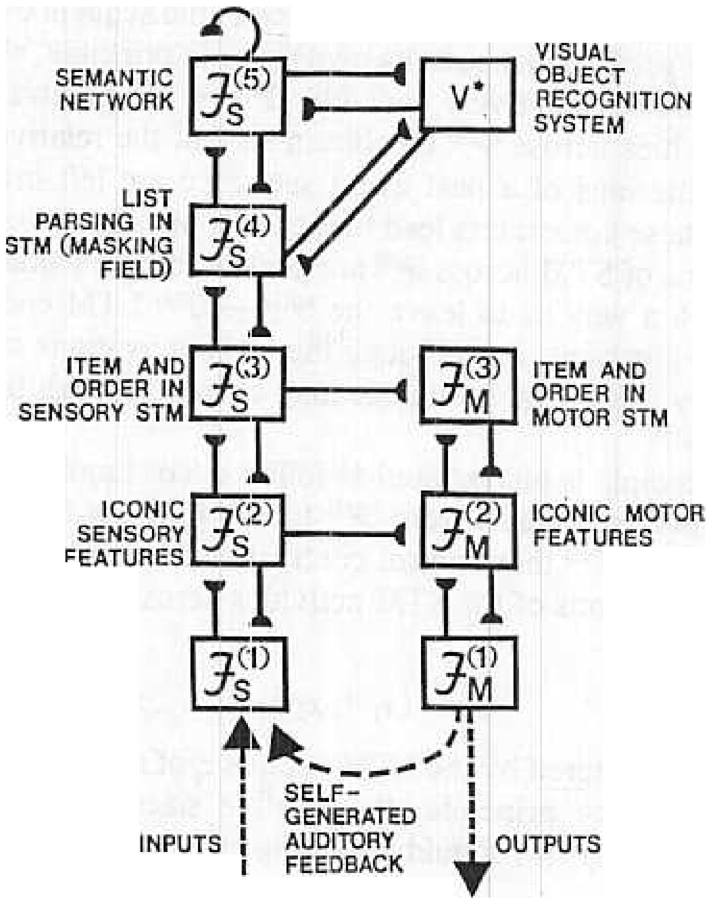

Figure 6.19 A macrocircuit governing self-organization of recognition and recall processes: Auditorily mediated language processes ($\mathcal{F}_S^{(i)}$), visual recognition processes (V*), and motor control processes ($\mathcal{F}_M^{(j)}$) interact internally via conditionable pathways (black lines) and externally via environmental feedback (dotted lines) to self-organize the various processes which occur at the different network stages.

self-organization approach to the unitization and temporal order problems that is invisible to a performance theoretic approach. It turns out that a shunting competitive network of a specialized design for $\mathcal{F}^{(3)}$ does the job.

Two considerations motivate this design. Once a sequence $r_1, r_2, \ldots, r_i$ has already been presented, its STM pattern represents "past" order information. Presenting a new item r_{i+1} can alter the total pattern of STM across $\mathcal{F}^{(3)}$, but I assume that this new STM pattern does not cause LTM recoding of that part of the pattern which represents past order information. New events are allowed to weaken the influence of codes that represent past order information but not to deny the fact that the past events occurred. This hypothesis prevents the LTM record of past order information from being obliterated by every future event that happens to occur.

This idea can be stated in a related way that emphasizes the possible destabilizing effects of new events when there are no homunculi present to beg the question. Every subsequence of the sequence $r_1, r_2, \ldots, r_i$ is a perfectly good sequence in its own right. In principle, all possible subsequences can be adaptively coded by $\mathcal{F}^{(3)} \rightarrow \mathcal{F}^{(4)}$ pathways. How can the STM activities across $\mathcal{F}^{(3)}$ be chosen so that the relative activities of all possible filterings of a past event sequence are left invariant by future events? These constraints lead to the *LTM invariance principle:* The spatial patterns of STM across $\mathcal{F}^{(3)}$ are generated by a sequentially presented list in such a way as to leave the $\mathcal{F}^{(3)} \rightarrow \mathcal{F}^{(4)}$ LTM codes of past event groupings invariant, even though the STM activations caused across $\mathcal{F}^{(3)}$ and $\mathcal{F}^{(4)}$ by these past groupings may change through time as new items activate $\mathcal{F}^{(3)}$.

This principle is instantiated as follows. To simplify the discussion, let the feedforward signals from $\mathcal{F}^{(3)}$ to $\mathcal{F}^{(4)}$ (but not the internal feedback signals within $\mathcal{F}^{(3)}$ that control contrast enhancement and normalization) be linear functions of the STM activities across $\mathcal{F}^{(3)}$. At time t_i, the STM pattern

$$P_i = (x_1^{(3)}, x_2^{(3)}, \ldots, x_i^{(3)}) \tag{46}$$

is adaptively filtered by the LTM vectors z_j of all nodes $v_j^{(4)}$ in $\mathcal{F}^{(4)}$. By the LTM invariance principle, the relative sizes of *all* the dot products $S_j(t_i) = P_i(t_i) \cdot z_j \cdot (t_i)$ should not change when r_{i+1} occurs. In other words,

$$P_i(t_{i+1}) \cdot z_j(t_i) = \omega_{i+1} P_i(t_i) \cdot z_j(t_i) \tag{47}$$

for all i and j, where ω_{i+1} is a proportionality constant that is independent of j. The LTM invariance principle thus implies that, after the STM traces $x_1^{(3)}, x_2^{(3)}, \ldots, x_i^{(3)}$ are excited by the items $r_1, r_2, \ldots, r_i$, they thereafter undergo proportional changes. The STM traces are shunted by multiplicative factors ω_{i+1}, ω_{i+2}, . . . that are independent of j.

Table 6.1 describes rules for generating these changes. In the table, the ith item r_i is instated in STM at $v_i^{(3)}$ with activity μ_i. At every successive item representation, all past STM traces are simultaneously shunted by the amounts ω_{i+1}, then ω_{i+2}, and so on. The STM activity of the ith item r_i after r_j occurs ($i < j$) is thus

$$x_i^{(3)}(t_j) = \mu_i \prod_{k=i+1}^{j} \omega_k .$$

It remains to be shown how the shunting parameters ω_k can be expressed in terms of the initial STM activities μ_i, where μ_i measures the amount of attention that is paid to r_i when it is first stored in STM. I accomplish this by using the fact that every shunting competitive network exhibits a normalization property to impose the following normalization rule.

The total STM activity across $\mathcal{F}^{(3)}$ after i items have been presented is

$$S_i = \mu_1 \phi_i + M(1 - \phi_i). \tag{49}$$

In Equation 49, ϕ_i is a positive decreasing function of i with $\phi_0 = 1$ and $\lim_{i\to\infty} \phi_i = 0$. By Equation 49, S_i grows from μ_1 to its asymptote $M(\geq \mu_1)$ as more items are stored. The *load parameters* ϕ_i estimate how close $\mathcal{F}^{(3)}$ is to saturating its total capacity. The load parameter ϕ_i estimate how close $\mathcal{F}^{(3)}$ is to saturating its total capacity. The load parameter ϕ_i also estimates how close $v_i^{(3)}$ is to the active item representations $v_1^{(3)}$, $v_2^{(3)}$, . . . , $v_{i-1}^{(3)}$ of previous events. A relatively large decrease of ϕ_i below ϕ_{i-1} means that $v_i^{(3)}$ gets activated with relatively little competition from previous items, due to the fact that r_i is represented in a different region of $\mathcal{F}^{(3)}$ than previous events. As more events are represented within $\mathcal{F}^{(3)}$, all regions of $\mathcal{F}^{(3)}$ become densely activated; hence $\lim_{i\to\infty} \phi_i = 0$. The special case $\phi_i = \theta^i$, where $0 < \theta < 1$, represents a field $\mathcal{F}^{(3)}$ whose activated item

Table 6.1
LTM Invariance Principle Constrains STM Activities of Sequentially Activated Item Representations

	$x_1^{(3)}$	$x_2^{(3)}$	$x_3^{(3)}$	$x_4^{(3)}$
$t \cong t_1$	μ_1	0	0	0
$t \cong t_2$	$\mu_1\omega_2$	μ_2	0	0
$t \cong t_3$	$\mu_1\omega_2\omega_3$	$\mu_2\omega_3$	μ_3	0

representations are uniformly spaced with respect to each other. In such a "homogeneous" field, $\phi_i/\phi_{i-1} = \theta$, which is independent of i.

By Equation 48, the total STM activity also equals

$$S_i = \sum_{j=1}^{i} \mu_j \prod_{k=j+1}^{i} \omega_k .$$

Equations 49 and 50 for S_i can be recursively identified to prove that the shunting weights satisfy the equation

$$\omega_k = \frac{S_k - \mu_k}{S_{k-1}} = \frac{\mu_1\phi_k + M(1 - \phi_k) - \mu_k}{\mu_1\phi_{k-1} + M(1 - \phi_{k-1})}$$

$k = 1, 2, \ldots$. By Equations 48 and 51,

$$x_i^{(3)}(t_j) = \mu_i \prod_{k=i+1}^{j} \left[\frac{\mu_1\phi_k + M(1 - \phi_k) - \mu_k}{\mu_1\phi_{k-1} + M(1 - \phi_{k-1})}\right].$$

Equation 52 characterizes STM across $\mathcal{F}^{(3)}$ for all time in terms of the attention paid to the items when they are stored (μ_i), the STM capacity of the network (M), and the load parameters (ϕ_i). Equation 52 can be rewritten in a way that suggests the relevance of probabilistic ideas to the STM temporal order problem. In terms of the notation $P_i(t_j) = x_i^{(3)}(t_j)S_j^{-1}$ and $p_i = \mu_i S_i^{-1}$, Equation 52 becomes

$$P_i(t_j) = p_i \prod_{k=i+1}^{j} (1 - p_k). \tag{53}$$

The STM patterns that evolve under the law in Equation 52 have been worked out in a number of cases (see Grossberg, 1978a, 1978e, Section 26). It is readily shown that a primacy effect often occurs in STM when a short subsequence of the list activates $\mathcal{F}^{(3)}$ but that this primacy effect is converted into an STM bow (primacy and recency effect) as more items are presented. For sufficiently long lists, the recency effect dominates the STM patterm. Multimodal bows, as in von Restorff STM effects, can also be generated under special circumstances.

All of the equations in this section have obvious generalizations to the case in which each item representation is distributed over many nodes. This is true because the shunting operations on the past field and the STM capacity of the network do not depend on how many nodes subserve each item representation. The same is true of the equations in the next section.

XXXV. TRANSIENT MEMORY SPAN, GROUPING, AND INTENSITY–TIME TRADEOFFS

Some remarks may help the reader to think about these STM results before I consider their implications for what is encoded in LTM. Given a fixed choice of the attentional sequence $\mu_1, \mu_2, \ldots$; the capacity M; and the inhibitory design $\phi_1, \phi_2, \ldots$, it follows that if $r_1, r_2, \ldots, r_K$ is the longest sublist that causes a primacy effect in STM, then every longer sublist $r_1, r_2, \ldots, r_K, r_{K+1}, \ldots, r_i$ will cause an STM bow at item r_K. I call K the *transient memory span* (TMS) of the list. The TMS is the longest sublist that can be directly read out of STM in the correct order. In Grossberg (1978e), I proved under weak conditions that the TMS is always shorter than the more familiar immediate memory span (IMS), which also benefits from LTM read-out. A typical choice of these parameters is TMS $\cong$ 4 and IMS $\cong$ 7 (Miller, 1956). One way to guarantee a correct read-out from STM without requiring template feedback from LTM is to rehearse the list items in subsequences, or groups, of a length no greater than the TMS.

In assigning the values μ_i and ω_k to the STM traces $x_i^{(3)}$, I have tacitly assumed that the times t_i at which items r_i are presented are sufficiently separated to enable these values to reach asymptote. If presentation rates are rapid, then only partial activations may occur, leading to weights of the form

$$\mu_i{}^* = \mu_i(1 - e^{-\lambda_i T_i}),$$

where λ_i is the rate of activation and $T_i = t_i - t_{i-1}$. Then the STM traces become

$$x_i^{(3)}(t_j) = \mu_i{}^* \prod_{k=i+1}^{j} \omega_k{}^*,$$

where by Equation 51,

$$\omega_k{}^* = \frac{\mu_1{}^*\phi_k + M(1 - \phi_k) - \mu_k{}^*}{\mu_1{}^*\phi_{k-1} + M(1 - \phi_{k-1})}.$$

Due to Equation 54, an intensity–time tradeoff, or Bloch's law (Repp, 1979), holds that may alter the STM pattern across $\mathcal{F}^{(3)}$ under conditions of rapid presentation. Such a tradeoff can limit the accuracy with which temporal order information is encoded in STM, most obviously by preventing some items from being stored in STM at all because they receive inadequate activation to exceed the network QT.

XXXVI. BACKWARD EFFECTS AND EFFECTS OF RATE ON RECALL ORDER

Two more subtle interactions of intensity and rate are worth noting. If items are rapidly presented but some are more drawn out than others, then the relative sizes of the STM activities can be changed. By changing the STM patterns across $\mathcal{F}^{(3)}$, the STM pattern across $\mathcal{F}^{(4)}$ that is caused by the adaptive filter $\mathcal{F}^{(3)} \rightarrow \mathcal{F}^{(4)}$ can also be changed. This STM pattern determines item recognition. Thus a change in rates can cause a contextually induced change in perception. In examples wherein items are built up from consonant and vowel sequences, a relative change in the duration of a later vowel may thus alter the perception of a prior consonant (Miller & Liberman, 1979; Schwab et al., 1981). Such examples support the hypothesis that STM patterns of temporal order information over item representation control network perception, not activations of individual nodes.

A uniform but rapid activation rate can alter both the items that are recalled and the order in which they are recalled. Suppose that the network is instructed to pay attention to a list during an attentional window of fixed duration. Whereas a slower presentation rate may allow a smaller number of items to be processed during this duration, a faster presentation rate may allow a larger number of items to be processed. In the former case, a primacy effect in STM may be encoded; hence correct read-out of order information from STM is anticipated. In the latter case, a bow in STM may be encoded. A fast rate may increase the number of items processed and thereby cause an STM bow in which items near the list middle are recalled worst (Grossberg, 1978a). A fast rate may also cause an STM bow in processing a fixed number of items if attention must be switched to the items as they are presented. Then the items near the list beginning and end may be recalled worst (Grossberg and Stone, 1986b; Reeves and Sperling, 1986; Sperling and Reeves, 1980).

XXXVII. SEEKING THE MOST PREDICTIVE REPRESENTATION: ALL LETTERS AND WORDS ARE LISTS

The LTM invariance principle indicates how a competitive shunting network can instate temporal order information in STM without destabilizing the LTM filters that learn from the STM patterns. We now ask how the outputs from all of these filters are interpreted at the next processing stage $\mathcal{F}^{(4)}$. How does $\mathcal{F}^{(4)}$ know which of its filtered inputs represent reliable data on which to base its output decisions? How does $\mathcal{F}^{(4)}$ select the codes for those sublists across $\mathcal{F}^{(3)}$ that are most predictive of the future? How does $\mathcal{F}^{(4)}$ know how to automatically group, or parse, the total event list represented across $\mathcal{F}^{(3)}$ into sublists that have the best a priori chance of predicting the future within the context defined by the unique past represented by that list?

The next few sections indicate how to design $\mathcal{F}^{(4)}$ so that its best predictive sublist chunks are assigned the greatest STM activity; how these most predictive chunks are differentially tuned by adaptive filtering and differentially gain control of predictive commands; how less predictive chunks are rapidly masked by more predictive chunks, therefore preventing the less predictive chunks from interfering with performance and enabling them to remain uncommitted by learning until they are unmasked in a different context where they are better predictors; how the masking due to predictive sublist chunks compresses the LTM code, computes a "magic number 7" (Miller, 1956), and changes the time scale of STM reset—and thus of LTM prediction—within the subfield of unmasked chunks; how the predictive recognition chunks remain uninhibited by rehearsal, since otherwise they could not sample the sequences to be learned and recalled; and how the predictive chunks can be directly inhibited only by other predictive chunks—say, those activated by new sensory feedback, or by nonspecific gain changes due to attention shifts.

The design of this field thus addresses the fundamental question of how "our conscious awareness . . . is driven to the highest level present in the stimulus" (Darwin, 1976). In contrast to the distinction made by McClelland and Rumelhart (1981) between a separate letter level and a word level (Section II), I suggest that "all letters and words are lists," indeed that all unitized events capable of being represented in $\mathcal{F}^{(4)}$ exist on an equal dynamical footing. This conclusion clarifies how changes in the context of a verbal item can significantly alter the processing of that item, and why the problem of identifying functional units has proved to be so perplexing (Studdert-Kennedy, 1980). In $\mathcal{F}^{(4)}$, no common verbal descriptor of the functional unit, such as phoneme or syllable or letter or word, has a privileged existence. Only the STM patterns of unitized chunks that survive the context-sensitive interaction between associative and competitive rules have a concrete existence. These rules instantiate principles of predictive stability that transcend the distinctions of lay language.

Before describing these rules, I should state what they do not imply. Despite the fact that "all letters and words are lists," a subject can be differentially set to respond to letters rather than words, numbers rather than letters, and so on. Such a capability involves the activation of learned top-down expectancies that selectively sensitize some internal representations more than others. Thus the phrase "all letters and words are lists" is a conclusion about the laws of unitization that letters and words share, not about the top-down attentional and expectational processes that can flexibly modulate the STM and LTM traces that these laws define.

XXXVIII. SPATIAL FREQUENCY ANALYSIS OF TEMPORAL PATTERNS BY A MASKING FIELD: WORD LENGTH AND SUPERIORITY

The main idea is to join together results about positional gradients, lateral masking, and multiple spatial frequency scales to synthesize an $\mathcal{F}^{(4)}$ network—called a *masking field*—that selectively amplifies the STM of $\mathcal{F}^{(4)}$ chunks representing longer sublists at the expense of chunks representing shorter sublists, other things being equal, up to some optimal sequence length (Grossberg, 1978e). Each of these three types of concepts can also be used to analyze aspects of spatial visual processing (Ganz, 1975; Robson, 1975). The network $\mathcal{F}^{(4)}$ thus illustrates that the same mechanisms can be specialized to do either spatial processing or temporal processing.

The results of Samuel et al. (1982, 1983) on a word length effect in word superiority studies were published after the first draft of this chapter was completed. That is, a letter is better recognized as it is embedded in longer words of lengths from 1 to 4. These authors write: "One could posit that the activation of a word is a function of the evidence present for it; more letters could provide more evidence for a word. Very short words would be at an inherent disadvantage, since they only receive a limited amount of support" (Samuel et al., 1983, p. 322). Both their data and their intuitive interpretation support the properties of word processing by a masking field developed in Grossberg (1978e). To clarify the critical issue of how "evidence" is defined to imply a word length effect, I have expanded my review of this issue in the subsequent sections, notably of how $\mathcal{F}^{(4)}$ chunks that represent longer item lists may mask $\mathcal{F}^{(4)}$ chunks that represent shorter item $\mathcal{F}^{(4)}$ lists. Several other predictions in Grossberg (1978e) have not yet been experimentally tested. Some of these predictions concern the rules of neuronal development whereby a masking field is self-organized. My expanded review forms a bridge between these and related levels of description.

XXXIX. THE TEMPORAL CHUNKING PROBLEM

The need for masking rules that are sensitive to the length of a list of items can be understood by considering the *temporal chunking problem:*

Suppose that an unfamiliar list of familiar items is sequentially presented (e.g., a novel word composed of familiar letters). In terms of frequency and familiarity, the most familiar units in the list are the items themselves. The first item starts to be processed before the whole list is even presented. What prevents processing of the first familiar item from blocking the chunking of the unfamiliar list? Another way to state this problem is as follows: In order to completely process a novel list, all of its individual items must first be presented. All of the items are more familiar than the list itself. What prevents item familiarity from forcing the list to always be processed as a sequence of individual items, rather than eventually as a unitized whole?

The temporal chunking problem is only recognized as a serious constraint on processing design when one analyzes frontally how wordlike representations are learned in response to serially scanned sound streams or visual letter arrays. To overcome this problem, somehow the sequence as a whole uses prewired processing biases to overcome, or mask, the learned salience of its constituent items.

The type of masking that I need goes beyond the usual masking models. To emphasize what is new, I briefly review some earlier masking models. The seminal model of Weisstein (1968, 1972) is a model of contrast enhancement. Ganz (1975) modified Weisstein's model to avoid its assumption that inhibition acts faster than excitation. Ganz's (1975) trace-decay-and-lateral-inhibition model is a special case of Equation 6. This model does not, however, discuss how the signal thresholds or LTM traces in Equation 6 interact with STM trace decay and lateral inhibition to alter a network's reaction time in response to target-then-mask. These factors were used by Grossberg (1969c) to provide a unified account of masking and performance speed-up due to learning. In this model, performance speed-up due to learning is a variant of the fan effect (Section II): An increase in a pathway's LTM trace amplifies signals in the pathway; these amplified signals more vigorously activate their receptive node; node activity therefore grows more rapidly and exceeds the output threshold of the node more quickly. The existence of more competing nodes can cause a larger total inhibitory signal to be received by each node; the net rate of growth of activity at each node is thereby decreased; node activity therefore exceeds the output threshold of the node less quickly. These properties also hold in the masking model that I now discuss.

This masking model is not just a model of contrast enhancement. It was introduced to analyze how developmental and attentional biases can alter competitive decision making before STM storage occurs (Grossberg & Levine, 1975). The model was extended to explain certain normative visual illusions, such as neutralization (Gibson, 1937; Levine & Grossberg, 1976). Both investigations analyzed how the STM decision process is altered by giving subsets of nodes different numbers of excitable sites, differentially amplified interaction strengths, and/or broader spatial interactions in shunting networks of the form

$$\frac{\mathrm{d}}{\mathrm{dt}} x_i = -Ax_i + (B_i - x_i)\left[I_i + \sum_{k=1}^{n} \mathrm{f}_k(x_k)C_{ki}\right] - (x_i + D)\left[J_i + \sum_{k=1}^{n} \mathrm{f}_k(x_k)E_{ki}\right], \tag{57}$$

$i = 1, 2, \ldots, n$, which are a special case of Equation 34. In these networks, a larger choice of coding sites B_i, or of shunting signals F_i in $\mathrm{f}_i(x_i) = \mathrm{f}(F_i x_i)$, or of spatial frequencies C_{ik} and E_{ik} endows a node v_i with the ability to mask the STM activities of nodes v_k with smaller parameters. The control of masking by parameters such as B_i, F_i, or E_{ik} is not the same process as contrast enhancement, since the latter is controlled by the choice of the signal function $\mathrm{f}(w)$ (Grossberg, 1973).

We discovered that a subtle interaction exists between the choice of parameters and signal functions. A linear signal function $\mathrm{f}(w)$ can cause the STM activities of all nodes with smaller parameters to be inhibited to zero no matter how big their STM activities start out relative to the STM activities of nodes with larger parameters. In such a network, structural or attentional biases (larger parameters) win out over the intensities or learned salience of individual cues (larger initial activities). This unsatisfactory state of affairs is overcome by using a sigmoid signal function $\mathrm{f}(w)$, in which case nodes with sufficiently large initial activities can mask nodes with larger parameters but smaller initial activities. Thus a flexible tug-of-war between stimulus factors, like intensity or learned salience, and structural factors, like the number of coding sites or spatial frequencies, exists if a nonlinear signal is used but not if a linear signal function is used. This fact poses yet another challenge to linear models.

Masking, as opposed to mere contrast enhancement, thus occurs in networks whose nodes are partitioned into subfields. Within each subfield, each node possesses (approximately) the same parameters. The interactions between nodes in different subfields are biased by the differences between subfield parameters.

XL. THE MASKING FIELD: JOINING TEMPORAL ORDER TO DIFFERENTIAL MASKING VIA AN ADAPTIVE FILTER

Although these masking insights were originally derived to study spatial processing in vision, I soon realized that they are useful, indeed crucial,

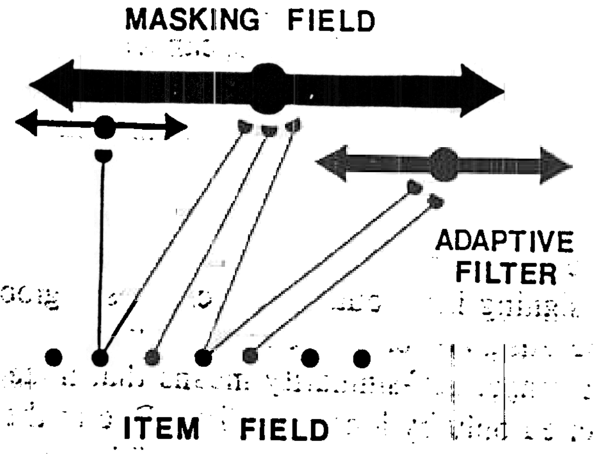

Figure 6.20 Selective activation of a masking field. The nodes in a masking field are organized so that longer item sequences, up to some optimal length, activate nodes with more potent masking properties. Individual items, as well as item sequences, are represented in the masking field. The text describes how the desired relationship between item field, masking field, and the intervening adaptive filter can be self-organized using surprisingly simple developmental rules.

for the study of temporal processing in language and motor control (Grossberg, 1978e). This realization came in stages. First I showed how the LTM invariance principle can be used to generate STM temporal order information over item representations (Section XXXIV). A spatial pattern of STM activity over a set of item representations in $\mathcal{F}^{(3)}$ encodes this information. As more items are presented, a new spatial pattern is registered that includes a larger region of the item field $\mathcal{F}^{(3)}$. The main insight is thus to translate the *temporal* processing of a list of items into a problem about a succession of expanding *spatial* patterns.

Given this insight, the temporal chunking problem can be translated as follows. How do chunks in $\mathcal{F}^{(4)}$ that encode broader regions of the item field $\mathcal{F}^{(3)}$ mask $\mathcal{F}^{(4)}$ chunks that encode narrower regions of $\mathcal{F}^{(3)}$? Phrased in this way, the relevance of the masking field results becomes obvious, because these results show how subfields with larger parameters can mask subfields with smaller parameters. Putting together these ideas about item coding and masking, the temporal chunking problem leads to the following design constraint: *Sequence masking principle:* Broader regions of the item field $\mathcal{F}^{(3)}$ are filtered in such a way that they selectively excite $\mathcal{F}^{(4)}$ nodes with larger parameters (Figure 6.20).

XLI. THE PRINCIPLE OF SELF-SIMILARITY AND THE MAGIC NUMBER 7

In order to realize this functional property in a computationally effective way, some specialized design problems must be solved. First, the masking parameters must be chosen self-consistently. It is inadmissible to allow a node v_i's larger number of sites B_i cancel the masking effect of its smaller spatial frequency E_{ik}. Nodes with more sites need to have broader interactions, other things being equal. This numerical constraint is a special case of a design principle that reappears in several guises throughout my work—the so-called *principle of self-similarity* (Grossberg, 1969e, 1982d). Every use of the principle suggests a different example wherein a local rule for designing individual cells achieves a global property that enhances the operating power of the entire network.

In the present usage, self-similarity means that nodes with larger parameters are excited only by longer sublists. Due to their self-consistent parameters, these nodes can be efectively inhibited only by other nodes that are also excited by longer sublists. Self-similarity thus introduces a pre-wired partial ordering among subfields such that the nodes activated by longer sublists can inhibit the nodes activated by shorter (and related) sublists, but not conversely, unless this partial ordering is modified through learning.

More list items need to be presented to activate a long-list node than a short-list node. Thus many more list items need to be presented to activate the same number of long-list nodes as short-list nodes. Since long-list nodes can be masked only by other long-list nodes, such nodes can remain active long enough to sample many more future events than can short-list nodes.

This last property shows how self-similarity enhances the network's predictive power. The network takes a risk by allowing any node to remain active for a long time. If the node samples inappropriate information on a given trial, on its next activation it can read out errors far into the future. This risk is minimized by letting the long-list nodes stay active the longest, because these nodes are better characterized by the temporal context (list length) into which they are embedded. Self-similarity is thus a structural constraint on individual nodes that enables the network as a whole to resolve uncertain input data without taking untoward predictive risks.

The abstract property of self-similarity also helps to explain a classic experimental property of human information processing, namely Miller's (1956) "magic number seven, plus or minus two." This is because the total length of the lists that can simultaneously be coded by a prescribed subfield increases with the total length of the sublists that can be chunked by the nodes of the subfield. Grossberg (1978e) discusses these properties in greater detail, notably their effects on word recognition, code compression, recall clustering effects, and the synthesis of predictive motor commands leading to rapid planned performance.

XLII. DEVELOPMENTAL EQUILIBRATION OF THE ADAPTIVE FILTER AND ITS TARGET MASKING FIELD

It remains for me to explain how the conditionable pathways that form the adaptive filter from the item field $\mathscr{F}^{(3)}$ to the masking field $\mathscr{F}^{(4)}$ generate the desired sublist masking properties. This explanation cannot merely offer formal rules for connecting the two fields. To be convincing, it must show how the connections can be established by growth rules that are simple enough to hold in vivo. The rules stated here thus amount to predictions about brain development in language-related anatomies.

The main properties to be achieved are all tacitly stated in the sequence masking principle. They may be broken down as follows:

1. *List representation*—The unordered sets of items in all realizable item lists, up to a maximal list length, are initially represented in the masking field.
2. *Masking parameters increase with list length*—The masking parameters of masking field nodes increase with the length of the item lists that activate them. This rule holds until an optimal list length is reached.
3. *Masking hierarchy*—A node that is activated by a given item list can mask nodes that are activated by sublists of this list.
4. *List selectivity*—If a node's trigger list has length n, it cannot be supraliminally activated by lists of length significantly less than n.

Properties 1 and 2 suggest that the adaptive filter contains a profusion of pathways that are scattered broadly over the masking field. Property 3 suggests that closely related lists activate nearby nodes in the masking field. Postulate 4 says that, despite the profusion of connections, long list nodes are tuned not to respond to short sublists.

The main problem is to resolve the design tension between profuse connections and list selectivity. This tension must be resolved both for short-list (e.g., letter) and long-list (e.g., word) nodes: If connections are profuse, why aren't short-list nodes unselective? In other words, what prevents many different item nodes from converging on every short-list node and thus being able to activate it? And if many item nodes do converge on long-list nodes, why aren't these long-list nodes activated by sublists of the items? Somehow the number of item nodes that contact a list node is calibrated to match the output threshold of the list node. A combination of random growth rules for pathways and self-similar growth rules for list nodes can be shown to achieve all of these properties (Cohen and Grossberg, 1986a; Grossberg, 1978e).

Suppose that each item node of $\mathscr{F}^{(3)}$ sends out a large number of randomly distributed pathways toward the list nodes of $\mathcal{F}^{(4)}$. Suppose further that an item node contacts a list node with a prescribed small probability p. This probability is small because there are many more list nodes than item nodes. Let λ be the mean number of such contacts across all of the list nodes. The probability that exactly k pathways contact a given list node is given by the Poisson distribution

$$P_k = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{58}$$

If K is chosen so that $K < \lambda < K + 1$, then P_k is an increasing function of k if $1 \leq k \leq K$ and a decreasing function of k if $k \geq K$. Thus lists of length k no greater than the optimal length K are represented within the masking field, thereby satisfying properties 1 and 2. Other random growth rules, such as the hypergeometric distribution, also have similar properties. Due to the broad and random distribution of pathways, list nodes will tend to be clustered near nodes corresponding to their sublists, thereby tending to satisfy property 3.

XLIII. THE SELF-SIMILAR GROWTH RULE AND THE OPPOSITES ATTRACT RULE

To discuss property 4, I interpret each list node as a population of cell sites. This population may consist of many neurons, each of which possesses many sites. For simplicity, I consider only a single neuron in such a population.

A list node that receives k pathways somehow dilutes the input due to each pathway so that (almost) all k pathways must be active to generate a suprathreshold response. As k increases, the amount of dilution also increases. This property suggests that long-list cells have larger cellular volumes, since a larger volume can more effectively dilute a signal due to a single output pathway. Larger volumes also permit more pathways to reach the cell's surface, other things being equal. The formal constraint that long-list nodes are associated with larger parameters, such as number of sites and spatial frequencies, is thus extended to a physical instantiation wherein more sites exist partly because the cells have larger surface areas. This conclusion reaffirms the importance of the self-similarity principle in designing a masking field: A cell has longer interactions (e.g., axons) because it has a larger cell body to support them.

How do larger cell surfaces attract more pathways, whereas smaller cell surfaces attract fewer pathways? This property is not as obvious as it may seem. Without further argument, a cell surface that is densely encrusted with axon terminals might easily be fired by a small subset of these axons. To avoid this possibility, the number of allowable pathways must be tuned so that the cell is never overloaded by excitation.

There exist two main ways to guarantee this condition. I favor the second way, but a combination of the two is also conceivable:

1. At an early stage of development, a spectrum of cell sizes is endogenously generated across the masking field by a developmental program. Each cell of a given size contains a proportional number of membrane organelles that can migrate and differentiate into mature membrane receptors in response to developing input pathways (Patterson & Purves, 1982). The number of membrane organelles is regulated to prevent the internal level of cell excitation (as measured, say, by the maximum ratio of free internal Na^+ to K^+ ions) from becoming too large.
2. Pathways from the item field grow to the list nodes via random growth rules. Due to random growth, some cells are contacted by more pathways than others. Before these pathways reach their target cells, these cells are of approximately the same size. As longer item lists begin to be processed by the item field, these lists activate their respective list nodes. The target cells experience an abnormal internal cellular milieu (e.g., abnormally high internal Na^+/K^+ concentration ratios) due to the convergence of many active pathways on the small cell volumes. These large internal signals gradually trigger self-similar cell growth that continues until the cell and its processes grow large enough to reduce the maximal internal signal to normal levels.

The tuning of cell volumes in $\mathcal{F}^{(4)}$ to the number of converging afferent pathways from $\mathcal{F}^{(3)}$ is thus mediated by a self-similar use-and-disuse growth rule. Grossberg (1969e) proposed such a rule for cell growth to satisfy general properties of cellular homeostasis. In the present application, the fact that internal cellular indices of membrane excitation can trigger cell growth until these indices equilibrate to normal levels suggests why the mature cell needs simultaneous activation from most of its pathways before it can fire.

A self-similar growth rule has many appealing properties. Most notably, only item lists that occur in a speaker's language during the critical growth period will be well represented by the chunks of the speaker's masking field. This fact may be relevant to properties of second language learning. If input excitation continues to maintain cell volume throughout the life of the cell, partial transaction of the cell's input pathways should induce a partial reduction in cell volume. Moreover, if the transient memory span of the item field equals K (Section XXXV), the optimal chunk length in the masking field should also approximate K. In other words, the chunk lengths (in the masking field $\mathcal{F}^{(4)}$) to which the speaker is sensitive are tuned by the lengths of item sequences (in the item field $\mathcal{F}^{(3)}$) that the speaker can recall directly out of STM.

A second issue concerning the developmental self-organization of the masking field is the following: How does each masking subfield know how to choose inhibitory pathways that are strong enough to carry out efficient masking but not so strong as to prevent any list from activating the masking field? I have predicted (Grossberg, 1978e, Section 45) that this type of property is developmentally controlled by an "opposites attract" rule, whereby excitatory sites attract inhibitory pathways and inhibitory sites attract excitatory pathways. The prediction suggests how intracellular parameters can regulate the attracting morphogens in such a way that balanced on-center, off-surround pathways result.

It remains to illustrate how constraints on list length, masking hierarchy, and list selectivity can be computationally realized in a network such as that in Equation 57. Cohen and Grossberg (1986a, 1986b) describe computer simulations that demonstrate all the desired properties of a masking field. This masking field obeys equations of the form

$$\frac{d}{dt} x_i^{(J)} = -Ax_i^{(J)} + (B - x_i^{(J)}) \left[\sum_{j \in J} I_j p_{ji}^{(J)} + D_{|J|} f(x_i^{(J)}) \right]$$
$$- (x_i^{(J)} + C) \sum_{m,K} g(x_m^{(K)}) F_{|J|} G_{|K|} H_{|K \cap J|}.$$

In Equation 59, $x_i^{(J)}$ is the STM activity of the ith cell in $\mathcal{F}^{(4)}$ which receives input pathways from only the item representations of items r_i, $i \in J$, where J is an unordered set of indices. Notation $|J|$ denotes the size of set J. Thus interaction coefficients such as $D_{|J|}$ and $F_{|J|}$ depend only upon the size of the set J of items, not upon the items themselves. These coefficients are chosen to satisfy the growth constraints

$$\sum_{j \in J} p_{ji}^{(J)} = \text{constant} \quad \text{and} \quad \sum_{m,K} F_{|J|} G_{|K|} H_{|K \cap J|} = \text{constant}.$$

XLIV. AUTOMATIC PARSING, LEARNED SUPERIORITY EFFECTS, AND SERIAL POSITION EFFECTS DURING PATTERN COMPLETION

I now summarize some of the psychological implications of masking field dynamics. As a list of items is presented to the item field $\mathcal{F}^{(3)}$, the encoding of the list will be updated continuously. At every time, the most predictive chunks of the list that is active at that moment will rapidly mask the activities of less predictive chunks, even though these less predictive chunks may have been dominant in an earlier temporal context. As the total list length exceeds the maximal length of any sublist encoded within the masking field $\mathcal{F}^{(4)}$, the network will automatically parse the total list into that grouping of sublists that can best survive mutual masking across $\mathcal{F}^{(4)}$.

Both of these properties are influenced by learning in important ways. For example, suppose that a given list is familiar to the network (e.g., a familiar word) but that none of its sublists has ever been individually presented to the network. In the $\mathcal{F}^{(3)} \rightarrow \mathcal{F}^{(4)}$ adaptive filter, the pattern of LTM traces corresponding to the chunk of the whole list will be much better tuned than the LTM patterns corresponding to any sublist chunk. This is because rapid masking of all sublist chunks by the whole list has occurred on all learning trials (Section XXIII). Consequently, on recall trials a sufficiently large sublist of the list may activate the chunk corresponding to the whole list rather than its own sublist chunk. This property is due to the differential amplification of the $\mathcal{F}^{(3)} \rightarrow \mathcal{F}^{(4)}$ signals that correspond to the tuned LTM traces of the list chunk. This whole-list pattern completion effect should be weaker in situations wherein the sublists are also familiar lists (e.g., familiar word embedded in familiar word) due to three factors working together: the stronger relative amplification of the sublist chunks by their tuned LTM patterns, the greater innate ease with which a sublist can activate a sublist chunk than a list chunk, and the possibility that nodes with smaller parameters can mask nodes with larger parameters if they receive larger inputs (Section XXXIX). These properties also indicate how a familiar word in a nonword may be recognized.

Which sublist of a list can best activate the full list code? Often the answer is the sublists concentrated at the beginning and end of a list. This is because the pattern of STM temporal order information across $\mathcal{F}^{(3)}$ often exhibits a primacy effect, a recency effect, or an STM bow (Section XXXIV). Thus the strongest STM activities are often at either end of a list. As the LTM pattern of a list chunk becomes parallel through learning to its STM pattern at $\mathcal{F}^{(3)}$, the largest LTM traces correspond to items at the list beginning and end. These large LTM traces are the ones capable of selectively amplifying sublist items. Rumelhart and McClelland (1982) report serial bowing effects in their data on word recognition. However, their model does not explain these data without the benefit of auxiliary hypotheses (see Lawry & LaBerge, 1981, for related data). If certain sublists are practiced often, their LTM patterns will preferentially activate the corresponding sublist chunks in the STM struggle across $\mathcal{F}^{(4)}$.

Different parsings of the same list can thus be determined by changing the alphabet of practiced sublists.

Once a given parsing of sublist codes across $\mathcal{F}^{(4)}$ starts to be activated, it delivers template feedback to $\mathcal{F}^{(3)}$. Word superiority effects (Johnston & McClelland, 1974) are abetted by the larger parameters of long lists, although there exists a tradeoff in reaction time measures of superiority between how long it takes to supraliminally activate a list code and the strength of its top-down feedback. The list length prediction has received experimental support from Samuel et al. (1982, 1983). A template explanation of word superiority is also suggested by Rumelhart and McClelland (1982). There are at least two important differences between our theories. As I noted in Section I, Rumelhart and McClelland (1982) postulate the existence of distinct letter and word levels, and connect letters and words to each other in different ways. My theory replaces letter and word levels by item and list levels, and connects these levels in a different way than would be appropriate if letter and word levels existed. These theories are fundamentally different both in their levels and in their interactions. For example, in a model using letter and word levels, letters such as A and I which are also words are represented on both levels, but letters such as K and L are represented only on the letter level. It remains unclear how such a distinction can be learned without using a homunculus. In contrast, in a model using item and list levels, all familiar letters are represented on both levels, because "all letters and words are lists." Although both letters and words can activate list chunks, they do so with varying degrees of ease due to differences in the spatial and temporal contexts into which they have been embedded.

This fact leads to a second major difference between our theories. Rumelhart and McClelland (1982) consider only four-letter words and do not discuss the role of learning. Therefore, they cannot easily explain how a familiar three-letter word in a four-letter nonword is processed. Instead of being able to use the *parametric biases* due to sublist length, learning, and so on in the coding of individual subsequences, they must derive all of the processing differences between words, pseudowords, and nonwords from differences in the number of activated words in their network hierarchy. This type of explanation does not seem capable of explaining the word length effect. Grossberg (1984b) and Grossberg and Stone (1986a) describe other differences between the theories and their ability to explain word recognition data.

The present theory suggests some of the operations that may prevent a word superiority effect from occurring (Chastain, 1982). Of particular interest is the manner in which attentional factors can modulate this effect. For example, suppose that a subject gives differential attention to the first item in a string, thus amplifying the corresponding item representation in STM. When the adaptive filter responds to the whole list of items, the input to the sublist code that corresponds to the first item is differentially amplified. The additional salience of this sublist code enables it to compete more effectively with the sublist codes of longer list chunks. This competition weakens the activation of the longer chunks in $\mathcal{F}^{(4)}$ and enables the item chunk to generate relatively more the template feedback to $\mathcal{F}^{(3)}$. Item chunk feedback is also the primary source of template feedback when a string of unrelated letters is presented. Attentional processes enter this explanation in two mechanistically distinct but interdependent ways. Attentional mechanisms amplify the item representation in $\mathcal{F}^{(3)}$. Template feedback is also an attentional mechanism but one that is capable of acting on a more global scale of processing. The two attentional mechanisms are linked via the adaptive filter and the masking field.

XLV. GRAY CHIPS OR GREAT SHIPS?

The resonant feedback dynamics between $\mathcal{F}^{(3)}$ and $\mathcal{F}^{(4)}$ also help to explain the interesting findings of Repp et al. (1978). By varying the fricative noise and silence durations in *gray ship*, they found that "given sufficient silence, listeners report GRAY CHIP when the noise is short but GREAT SHIP when it is long" (Repp et al., 1978, p. 621). There exists "a trading relation between silence and noise durations. As noise increases more silence is needed . . . For equivalent noise durations, more silence was needed in the fast sentence frame than in the slow sentence frame to convert the fricative into an affricative" (Repp et al., 1978, p. 625).

Part of an explanation for this phenomenon depends on the fact that articulatory acts influence which "feature detectors" are tuned by auditory feedback (Section XXX). Auditory experience of articulatory acts thus determines not only what item representations of $\mathcal{F}^{(3)}$ will be activated, but also what sequence representations of $\mathcal{F}^{(4)}$ will be activated. Another part of the explanation uses the fact that the list codes in $\mathcal{F}^{(4)}$ group together and perceptually complete auditory signals into familiar articulatory configurations via learned template feedback to $\mathcal{F}^{(3)}$. A subtle issue here is that a particular completion by template feedback is often contingent on the receipt of at least partially confirmatory auditory cues. Yet another part of the explanation uses the fact that a speed-up of speaking rate may alter commensurately all STM activities across $\mathcal{F}^{(3)}$ by changing all the item integration times, as in Equation 54. Due to the LTM invariance principle, a sufficiently uniform speed-up may not significantly alter the list codes selected across $\mathcal{F}^{(4)}$ (Section XXXV) after constrast enhancement has acted to generate tuned categories (Section XXII). Thus "judgments of phonetic structure and tempo are not independent, but are made simultaneously and interactively" (Miller, 1981, p. 69).

Finally, we come to the role of silence, which I consider the most challenging aspect of these data. Silence is not a passive state of "nothingness"; it is an active state that reflects the temporal context in which it is placed. Apart from the featural properties of silence as a temporal boundary to activity pattern onsets and offsets, I believe that the trading relationship reflects the fact that the nonspecific gain of $\mathcal{F}^{(4)}$ is higher during rapid speech than during slower speech and that this gain varies on a slower time scale than the onset or offset of an individual auditory cue. Recall from Section XXIII that a nonspecific gain control signal accompanies each specific cue to regulate the network QT or, equivalently, to renormalize the total operating load on the network (see also Grossberg, 1978e, Section 59). Such a variation of gain with speech rate can partially compensate for a decrease in integration times T_i by increasing λ_i in Equation 54 and decreasing ϕ_k in Equation 56. Thus the effects of a given duration of silence can be interpreted only by knowing the context-sensitive gain that calibrates the processing rates of auditory cues which bound the silence. These properties need to be studied further in numerical simulations.

XLVI. SENSORY RECOGNITION VERSUS MOTOR RECALL: NETWORK LESIONS AND AMNESIAS

This chapter's summary of the temporal coding designs that are presently known is incomplete. One also needs to build up analogous machinery to chunk the temporal order of motor commands and then to show how sensory and motor chunks are interconnected via associative maps. Only in this fashion can a full understanding of the differences and relationships between sensory recognition and motor recall be achieved.

The partial independence of sensory and motor temporal order mechanisms is perhaps best shown through the behavior of amnesic patients (Butters & Squire, 1983). Formal amnesic syndromes that are strikingly reminiscent of real amnesic syndromes can be generated in the networks of the present theory. For example, cutting out the source of orienting arousal in Figure 6.15 generates a network that shares many symptoms of medial temporal amnesia, and cutting out the source of incentive motivational feedback in network models of motivated behavior generates a syndrome characterized by flat affect and impaired transfer from sensory

STM to LTM (Grossberg, 1971, 1975, 1982b). Both of these lesions are interpreted as occurring in formal network analogs of hippocampus and other closely related structures.

Grossberg (1978e, Section 34) analyzes a network (Figure 6.19) in which sensory and motor temporal coding mechanisms are associatively joined to allow updating of internal representations to take place during the learning and performance of planned action sequences. The next section supplements these designs by outlining some related concepts about rhythm.

XLVII. FOUR TYPES OF RHYTHM: THEIR REACTION TIMES AND AROUSAL SOURCES

I believe that humans possess at least four mechanistically distinct sources of rhythmic capability. The on–off rebounds within specialized gated dipole circuits can be used to generate endogenous rhythms (Carpenter and Grossberg, 1983a, 1983b, 1984, 1985), as in the periodic rhythms of agonist–antagonistic motor contractions. For example, suppose that the on-cell of a gated dipole is perturbed to get a rhythm started. A few controlled on-cell inputs at a fixed rate can determine the nonspecific arousal level (Figure 6.21), which then feeds back to maintain an automatic on–off oscillation at the same rate, as in walking, until the arousal level is inhibited. Willed changes in the arousal level can continuously modulate the frequency of the oscillation after it gets going. When successive dipole fields in a network hierarchy interact mutually, they can also mutually entrain one another in a rhythmic fashion (Grossberg, 1978f). A deep understanding of this type of entrainment requires further numerical and mathematical study of the nonlinear dynamics of interacting dipole fields.

In a related type of rhythm generator, source cells excite themselves with positive feedback signals and inhibit other source cells via inhibitory interneurons that temporally average outputs from the source cells. In these on-center off-surround networks, the temporal averaging by inhibitory interneurons replaces the temporal averaging by transmitter gates which occur in a gated dipole. A nonspecific arousal signal which equally excites all the source cells energizes the rhythm and acts as a velocity signal. Ellias and Grossberg (1975) showed that in-phase oscillations can occur when the arousal level is relatively small. As the arousal level is increased, these in-phase oscillations occur with higher frequency. When a critical arousal level is reached, a Hopf bifurcation occurs. Out-of-phase oscillations then occur at increasingly high frequency as the arousal level

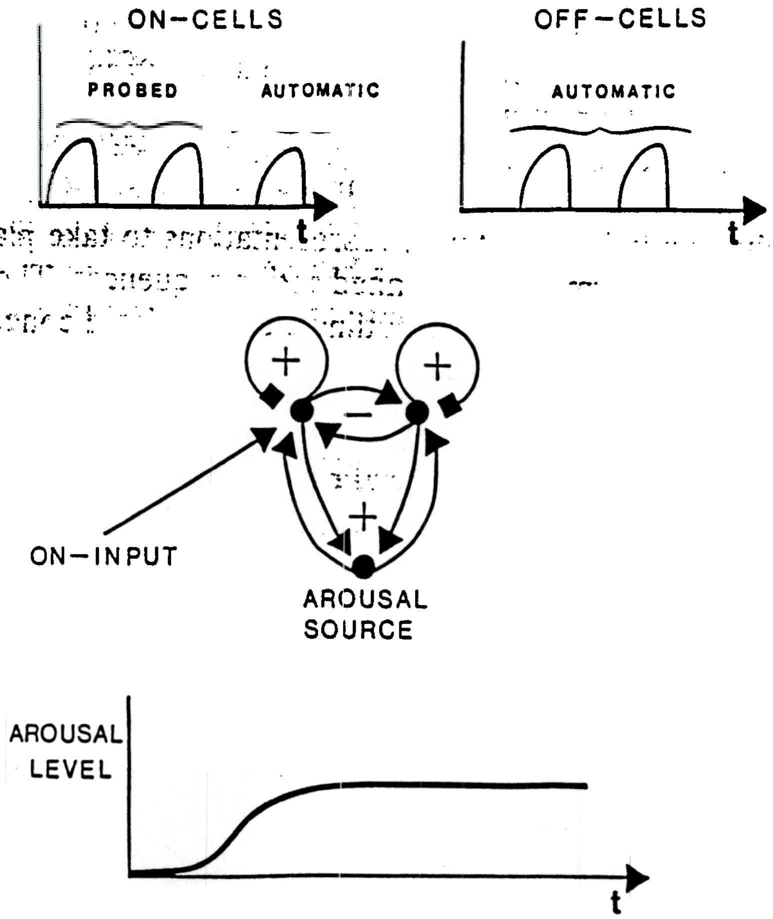

Figure 6.21 A feedback gated dipole as a rhythm generator: A few evenly spaced on-inputs to the gated dipole start an out-of-phase oscillation going between on-cells and off-cells. Both types of cells can activate the nonspecific arousal node which has been sensitized by an act of will. The period of the rhythm sets the average level of arousal, which in turn perpetuates the rhythm until the arousal node is inhibited or further excited to speed up the rhythm.

is further increased. Such results suggest an approach towards understanding how changes in motor gait are controlled by spinal circuits which automatically interpret a simple descending velocity signal by changing both the frequency and the patterning of motor outflow signals to several limbs (Grillner, 1975).

A third type of rhythm occurs when a preplanned "program" of actions is read out of a pattern of temporal order information in STM as rapidly as possible (Section XVI). This type of rhythm also uses nonspecific arousal, in the form of a rehearsal wave, abetted by self-inhibitory feedback that sequentially resets the STM pattern to prevent a single action from being performed perseveratively. Performance of this kind can exhibit at least two properties: an increase in the reaction time of the first item as a function of list length, due to normalization of the total STM activity—

Section XVIII, and a slowing down of the performance of later items, due to the tendency for primacy to dominate recency in short lists (Sections VIII and XXXIV). Sternberg and his colleagues have reported reaction time data of this type during rapid speaking and typewriting of word lists of different lengths (Sternberg et al., 1978; Sternberg et al., 1980). Grossberg and Kuperstein (1986) have used this type of rhythm generator to analyze how a sequence of planned eye movements can be performed under control of the frontal eye fields.

I call the fourth type of rhythmic capability "imitative rhythm." This is the type of rhythm whereby a list of familiar items of reasonable length can be performed at a prescribed *aperiodic* rhythm after a single hearing. It is also the type of rhythm whereby one can think of DA-DA not as a list of four symbols but as DA repeated twice. This example suggests the relevance of interactions between ordered item representations and a rhythm-generating mechanism to the development of simple counting skills (Gelman & Gallistel, 1978). The mechanism of imitative rhythms also uses a nonspecific arousal mechanism. Indeed, all rhythmic mechanisms use nonspecific arousal in some way, and all nonspecific arousal sources elicit rhythm by being interpreted by the field of specific representations that they energize (Section IX). In this sense, all rhythmic mechanisms are structural expressions of the factorization of pattern and energy (Section V).

The intuitive idea leading to a mechanism of imitative rhythm is depicted in Figure 6.22. Each list item is encoded by adaptive filtering and a temporal order representation in STM. Each list item also simultaneously delivers a nonspecific arousal pulse to the rhythm generator. Thus every item excites a specific pathway and a nonspecific pathway in a variant of Figure 6.15. The internal organization of the rhythm generator converts the *duration* of the nonspecific pulse into a topographically organized STM *intensity*. This happens for every item in the sequence, up to some capacity limit. Thus the rhythm generator is a (parallel) buffer of sorts, but it does not encode item information. Rather, it codes rhythm abstractly as an ordered series of intensities. When the rehearsal wave nonspecifically activates the whole field, these intensities are read out in order by a parallel mechanism and are reconverted into durations. A detailed construction of a rhythm generator is given in Grossberg (1985).

Coding a series of durations as a spatially ordered pattern of STM intensities greatly simplifies the efficient learning of aperiodic sequences of actions. For example, a single sequence chunk in $\mathcal{F}^{(4)}$ can simultaneously sample a spatial pattern of temporal order information in STM over a field of item representations *and* a spatial pattern of ordered intensities in the rhythm generator. Read-out from a single sequence chunk can

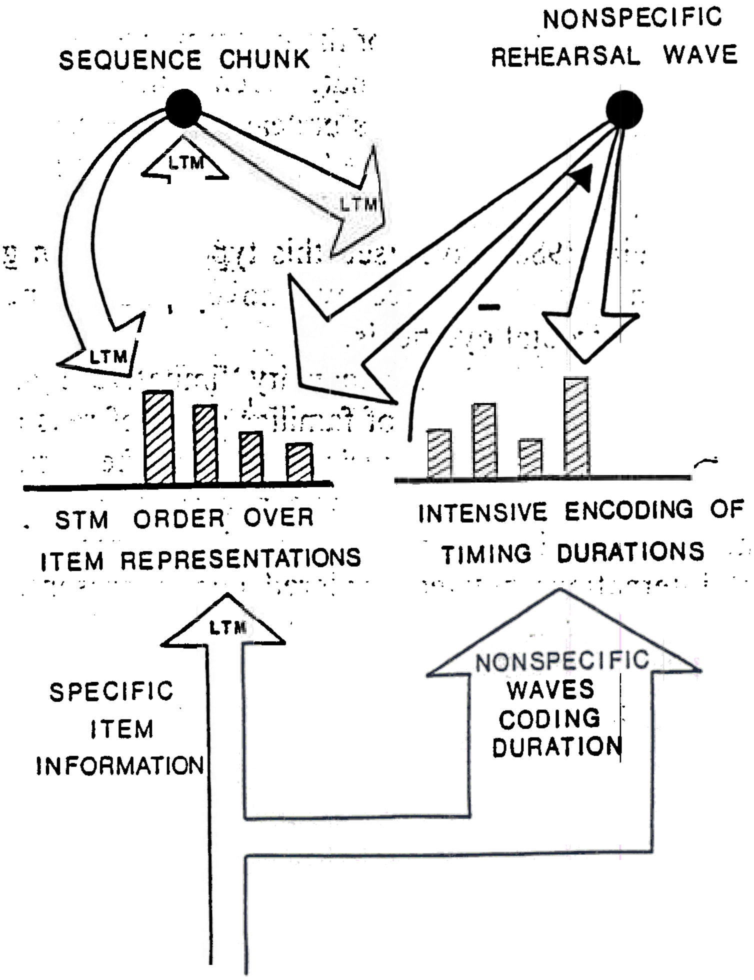
SEQUENCE CHUNK
NONSPECIFIC REHEARSAL WAVE
LTM
LTM
LTM
LTM
−
STM ORDER OVER ITEM REPRESENTATIONS
INTENSIVE ENCODING OF TIMING DURATIONS
SPECIFIC ITEM INFORMATION
NONSPECIFIC WAVES CODING DURATION

Figure 6.22 An aperiodic rhythm generator transforms durations of nonspecific arousal waves into ordered intensities of stored STM activities. Concurrently, the list items that elicited the arousing signals are stored as a pattern of STM temporal order information across item representations. The onset of a sustained nonspecific rehearsal wave transforms the ordered intensities of the rhythm generator back into durations during which the rehearsal wave is inhibited. The result is a series of timed and ordered output bursts from the item representations. Rapid performance of the item representations can occur in response to the rehearsal wave alone even when no inhibiting signals from the rhythm generator are available to modulate performance rate. Both the STM pattern of temporal order information and the STM pattern of timing information can be encoded by a single node, such as the sequence chunk that is generated by adaptive filtering of the STM pattern of ordered item information.

thus recall the entire sequence of items with the correct aperiodic rhythm. For example, a baby may repeat the correct number of sounds in response to an unfamiliar list of words, as well as the rhythm with which the sounds were spoken, even though he or she cannot pronounce the sounds themselves. The ability to imitate the rhythm of one, two, or three sounds is at first much better than the ability to imitate the rhythm or number of a longer list of sounds.

XLVIII. CONCLUDING REMARKS

This chapter illustrates how a small number of network principles and mechanisms can be used to discuss many topics related to the adaptive self-organization of serial order in behavior. Perhaps the most important unifying concepts that arise in this framework are those of adaptive resonance, adaptive context-mediated avalanche, adaptively invariant STM order information in an item field, and an adaptively tuned self-similar masking field. All of these concepts suggest that the functional units of network activity are inherently nonlinear and nonlocal patterns that coherently bind a network's local computations into a context-sensitive whole. The program of classifying the adaptive resonances that control different types of planned serial behavior promises to antiquate the homunculi that burden some contemporary theories of intelligent behavior, and to end Neisser's (1976) nightmare of "processing and still more processing" with a synthetic moment of resonant recognition.

APPENDIX: DYNAMICAL EQUATIONS

A few equations include all the constructions of embedding field theory. Although it is hard work to choose the parameters that characterize specialized processors, these equations provide a guiding framework.

When Equation 6 is generalized to include conditionable inhibitory LTM traces z_{ij}^- as well as conditionable excitatory LTM traces z_{ij}^+, we find (using an obvious extension of the notation) that

$$\frac{d}{dt} x_i = -A_i x_i + \sum_j B_{ji} z_{ji}^+ - \sum_i C_{ji} z_{ji}^- + I_i,$$

$$\frac{d}{dt} z_{ij}^+ = -D_{ij}^+ z_{ij}^+ + E_{ij}^+ [x_j]^+$$

and

$$\frac{d}{dt} z_{ij} = -D_{ij}^- z_{ij}^- + E_{ij}^- [x_j]^-,$$

where $[\xi]^- = \max(-\xi, 0)$. If the inhibitory signals are mediated by slowly varying inhibitory interneuronal potentials x_i^- that are activated by excitatory potentials x_i^+, we find that

$$\frac{d}{dt} x_i^+ = -A_i^+ x_i^+ + \sum_i B_{ji}^{++} z_{ji}^{++} - \sum_i C_{ji}^{-+} z_{ji}^{-+} + I_i^{++}$$

$$\frac{d}{dt} x_i^- = -A_i^- x_i^- + \sum_j B_{ji}^{+-} z_{ji}^{+-} - \sum_j C_{ji}^{--} z_{ji}^{--} + I_i^{--}, \qquad \text{(A5)}$$

In Equation A4, B_{ji}^{++} denotes an excitatory signal from v_j^+ to v_i^+ and C_{ji}^{-+} denotes an inhibitory signal from v_j^- to v_i^+. The other notations can be read analogously. Four types of LTM traces are now possible; for example,

$$\frac{d}{dt} z_{ij}^{-+} = -D_{ij}^{-+} z_{ij}^{-+} + E_{ij}^{-+}[x_j]^+, \qquad \text{(A6)}$$

and

$$\frac{d}{dt} z_{ij}^{--} = -D_{ij}^{--} z_{ij}^{--} + E_{ij}^{--}[x_j]^-. \qquad \text{(A7)}$$

If interactions can be either shunting or additive, then equation A4 is generalized to

$$\frac{d}{dt} x_i^+ = -A_i^+ x_i^+ + (F_i^+ - G_i^+ x_i^+)\left[\sum_j B_{ji}^{++} z_{ji}^{++} + I_i^{++}\right]$$
$$(H_i^+ + K_i^+ x_i^+)\left[\sum_j C_{ji}^{-+} z_{ji}^{-+} + J^{-+}\right] \qquad \text{(A8)}$$

Equations A5–A7 have similar generalizations. For example, Equation A6 becomes

$$\frac{d}{dt} z_{ij}^{-+} = -D_{ij}^{-+} z_{ij}^{-+} + (L_{ij}^{-+} - M_{ij}^{-+} z_{ij}^{-+}) E_{ij}^{-+}[x_j]^+ \qquad \text{(A9)}$$

If transmitter accumulation rate is slow relative to transmitter depletion rate, then the amount of transmitter Z_{ij}^{-+} generated by the LTM trace z_{ij}^{-+} satisfies

$$\frac{d}{dt} Z_{ij}^{-+} = (N_{ij}^{-+} z_{ij}^{-+} - P_{ij}^{-+} Z_{ij}^{-+}) - Q_{ij}^{-+} Z_{ij}^{-+} \qquad \text{(A10)}$$

where Q_{ij}^{-+} increases with x_i^-. The transmitter gating equations of a gated dipole are of this type. Correspondingly, Equation A8 is changed to

$$\frac{d}{dt} x_i^+ = -A_i^+ x_i^+ + (F_i^+ - G_i^+ x_i^+)\left[\sum_j B_{ji}^{++} Z_{ji}^{++} + I_i^{++}\right]$$
$$- (H_i^+ + K_i^+ x_i^+)\left[\sum_j C_{ji}^{-+} Z_{ji}^{-+} + J_i^{-+}\right] \qquad \text{(A11)}$$

If self-regulatory autoreceptive feedback occurs among all the synapses of similar type that converge on a single node, then Equation A10 becomes

$$\frac{d}{dt} Z_{ij}^{-+} = (N_{ij}^{-+} z_{ij}^{-+} - P_{ij}^{-+} Z_{ij}^{-+}) - \sum_k R_{kj}^{+} Q_{kj}^{-+} Z_{kj}^{-+}. \qquad \text{(A12)}$$

The other transmitter equations admit analogous autoreceptor generalizations. Transient properties of transmitters, such as mobilization and enzymatic modulation, may be defined by extensions of these equations (Carpenter & Grossberg, 1981; Grossberg, 1974).

REFERENCES

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406.

Anderson, J. R. (1983). Retrieveal of information from long-term memory. *Science, 220,* 25–30.

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

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Advances in the phychology of learning and motivation research and theory* (Vol. 2). New York: Academic Press.

Banquet, J.-P., & Grossberg, S. (1986). Structure of event-related potentials during learning: An experimental and theoretical analysis. Submitted for publication.

Berger, T. W., & Thompson, R. F. (1978). Neuronal plasticity in the limbic system during classical conditioning of the rabbit nictitating membrane response, I: The hippocampus. *Brain Research, 145,* 323–346.

Butters, N., & Squire, L. (Eds.). (1983). *Neuropsychology of memory.* New York: Guilford Press.

Carney. A. E., Widen, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America, 62,* 961–970.

Carpenter, G. A., & Grossberg, S. (1981). Adaptation and transmitter gating in vertebrate photoreceptors. *Journal of Theoretical Neurobiology, 1,* 1–42.

Carpenter, G. A., & Grossberg, S. (1983a). A neural theory of circadian rhythms: The gated pacemaker. *Biological Cybernetics, 48,* 35–59.

Carpenter, G. A., & Grossberg, S (1983b). Dynamic models of neural systems: Propagated signals, photoreceptor transduction, and circadian rhythms. In R. Grissell, J.P.E. Hodgson, & M. Yanowich (Eds.), *Oscillations in mathematical biology.* New York: Springer-Verlag.

Carpenter, G. A., & Grossberg, S. (1984). A neural theory of circadian rhythms: Aschoff's rule in diurnal and nocturnal mammals. *American Journal of Physiology, 247,* R1067–R1082.

Carpenter, G. A., & 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., & Grossberg, S. (1986a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing,* in press.

Carpenter, G. A., & Grossberg, S.(1986b). Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In J. Davis, R. Newburgh, & E. Wegman (Eds.), *Brain structure, learning, and memory.* AAAS Symposium Series, in press.

Carpenter, G. A., & Grossberg, S. (1986c). Neural dynamics of category learning and recognition: Structural invariants, reinforcement, and evoked potentials. In M. L. Commons, S. M. Kosslyn, and R. J. Herrnstein (Eds.), *Pattern recognition and concepts in animals, people, and machines*. Hillsdale, NJ: Erlbaum.

Cermak, L. S., & Craik, F.I.M. (1979). *Levels of processing in human memory*. Hillsdale, NJ: Erlbaum.

Chastain, G. (1982). Scanning, holistic encoding, and the word-superiority effect. *Memory and Cognition, 10,* 232–236.

Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage in competitive neural networks. IEEE *Transactions, smc13,* 815–826.

Cohen, M. A., & Grossberg, S. (1986a). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, in press.

Cohen, M. A., & Grossberg, S. (1986b). Unitized recognition codes for parts and wholes: The unique cue in configural discriminations. In M. L. Commons, S. M. Kosslyn, & R. J. Herrnstein (Eds.), *Pattern recognition and concepts in animals, people, and machines*. Hillsdale, NJ: Erlbaum.

Cole, R. A., Rudnicky, A. I., Zue, V. W., & Reddy, D. R. (1980). Speech as patterns on paper. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic memory. *Psychological Review, 82,* 407–428.

Cooper, W. E. (1979). *Speech perception and production: Studies in selective adaptation*. Norwood, NJ: Ablex.

Cornsweet, T. N. (1970). *Visual perception*. New York: Academic Press.

Crowder, R. G. (1978). Mechanisms of auditory backward masking in the stimulus suffix effect. *Psychological Review, 85,* 502–524.

Dallet, K. M. (1965). "Primary memory": The effects on redundancy upon digit repetition. *Psychonomic Science, 3,* 365–373.

Darwin, C. J. (1976). The perception of speech. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. VII), New York: Academic Press.

Deadwyler, S. A., West, M. O., & Robinson, J. H. (1981). Entorhinal and septal inputs differentially control sensory-evoked responses in the rat dentate gyrus. *Science, 211,* 1181–1183.

DeFrance, J. F. (1976). *The septal nuclei*. New York: Plenum Press.

Dethier, V. G. (1968). *Physiology of insect senses*. London: Methuen.

Dixon, T. R., & Horton, D. L. (1968). *Verbal behavior and general behavior theory*. Englewood Cliffs, NJ: Prentice-Hall.

Dodwell, P. C. (1975). Pattern and object perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. V). New York: Academic Press.

Eccles, J. C. (1952). *The neurophysiological basis of mind: The principles of neurophysiology*. London: Oxford University Press.

Eccles, J. C., Ito, M., & Szentagothai, J. (1967). *The cerebellum as a neuronal machine*. New York: Springer-Verlag.

Ellias, S. A., & Grossberg, S. (1975). Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks. *Biological Cybernetics, 20,* 69–98.

Elman, J. H., Diehl, R. L., & Buchwald, S. E. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America, 62,* 971–974.

Erickson, R. P (1963). Sensory neural patterns and gustation. In Y. Zotterman (Ed.), *Olfaction and taste*. New York: Pergamon Press.

Estes, W. K. (1972). As associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory*. New York: John Wiley.

Fisher, R. P., & Craik, E.I.M. (1980). The effects of elaboration on recognition memory. *Memory & Cognition, 8*, 400–404.

Fitts, P. M., & Posner, M. L. (1967). *Human performance*. Monterey, CA: Brooks/Cole.

Foss, D. J., & Blank, M. A. (1980). Identifying the speech codes. *Cognitive Psychology, 12*, 1–31.

Fowler, C. A. (1977). *Timing control in speech production*. Unpublished doctoral dissertation, Dartmouth College, Hanover NH.

Freeman, W. J. (1975). *Mass action in the nervous system*. New York: Academic Press.

Freeman, W. J. (1979). EEG analysis gives models of neuronal template-matching mechanism for sensory search with olfactory bulb. *Biological Cybernetics, 35*, 221–234.

Fry, D. B. (1966). The development of the phonological system in the normal and the deaf child. In F. Smith & G. A. Miller (Eds.), *The genesis of language*. Cambridge, MA: MIT Press.

Fukushima, K. (1980). Neocognition: A self-organized neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Gabriel, M., Foster, K., Orona, E., Saltwick, S. E., & Stanton, M. (1980). Neuronal activity of cingulate cortex, anteroventral thalamus, and hippocampal formation in discrimination conditioning: Encoding and extraction of the significance of conditional stimuli. *Progress in Psychobiology and Physiological Psychology, 9*, 125–231.

Ganz, L. (1975). Temporal factors in visual perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. V). New York: Academic Press.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gibson, J. J. (1937). Adaptation, after-effect and contrast in the perception of tilted lines, II: Simultaneous contrast and the areal restriction of the after-effect. *Journal of Experimental Psychology, 20*, 553–569.

Grillner, S. (1975). Locomotion in vertebrates: Central mechanisms and reflex interaction. *Physiological Review, 55*, 247–304.

Grossberg, S. (1964). *The theory of embedding fields with applications to psychology and neurophysiology*. New York: Rockefeller Institute for Medical Research.

Grossberg, S. (1967). Nonlinear difference-differential equations in prediction and learning theory. *Proceedings of the National Academy of Science, 58*, 1329–1334.

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

Grossberg, S. (1968b). Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Science, 60*, 758–765.

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

Grossberg, S. (1969b). On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics, 1*, 319–350.

Grossberg, S. (1969c). On learning, information, lateral inhibition, and transmitters. *Mathematical Biosciences, 4*, 255–310.

Grossberg, S. (1969d). On learning of spatiotemporal patterns by networks with ordered sensory and motor components: Excitatory components of the cerebellum. *Studies in Applied Mathematics, 48,* 105–132.

Grossberg, S. (1969e). On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology, 22,* 325–264.

Grossberg, S. (1969f). On the serial learning of lists. *Mathematical Biosciences, 4,* 201–253.

Grossberg, S. (1969g). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I. *Journal of Mathematics and Mechanics, 19,* 53–91.

Grossberg, S. (1970a). Neural pattern discrimination. *Journal of Theoretical Biology, 27,* 291–337.

Grossberg, S. (1970b). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Mathematics, 49,* 135–166.

Grossberg, S. (1971). On the dynamics of operant conditioning. *Journal of Theoretical Biology, 33,* 225–255.

Grossberg, S. (1972a). A neural theory of punishment and avoidance, I: Qualitative theory. *Mathematical Biosciences, 15,* 39–67.

Grossberg, S. (1972b). A neural theory of punishment and avoidance, II Quantitative theory. *Mathematical Biosciences, 15,* 253–285.

Grossberg, S. (1972c). Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt (Eds.), *Delay and functional-differential equations and their applications.* New York: Academic Press.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52,* 217–257.

Grossberg, S. (1974). Classical and instrumental learning by neural networks. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (Vol. 3). New York: Academic Press.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology, 18,* 263–327.

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

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics, 23,* 187–202.

Grossberg, S. (1977). Pattern formation by the global limits of a nonlinear competitive interaction in *n* dimensions. *Journal of Mathematical Biology, 4,* 237–256.

Grossberg, S. (1978a). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology, 17,* 199–219.

Grossberg, S. (1978b). Communication, memory, and development. In R. Rosen & F. Snell (Eds.), *Progress in Theoretical Biology* (Vol. 5). New York: Academic Press.

Grossberg, S. (1978c). Decisions, patterns, and oscillations in the dynamics of competitive systems with applications to Volterra-Lotka systems. *Journal of Theoretical Biology, 73,* 101–130.

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

Grossberg, S. (1978e). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (Vol. 5). New York: Academic Press.

Grossberg, S. (1978f). A theory of visual coding, memory, and development. In E. Leeuwenberg & H. Buffart (Eds.), *Formal theories of visual perception.* New York: John Wiley.

Grossberg, S. (1980a). Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences, 77,* 2338–2342.

Grossberg, S. (1980b). Direct perception or adaptive resonance? *Behavioral and Brain Sciences, 3,* 385.

Grossberg, S. (1980c). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

Grossberg, S. (1980d). Human and computer rules and representations are not equivalent. *Behavioral and Brain Sciences, 3,* 136–138.

Grossberg, S. (1981a). Adaptive resonance in development, perception, and cognition. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology.* Providence, RI: American Mathematical Society.

Grossberg, S. (1981b). Psychophysiological substrates of schedule interactions and behavioral contrast. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology.* Providence, RI: American Mathematical Society.

Grossberg, S. (1982a). Associative and competitive principles of learning and development: The temporal unfolding and stability of STM and LTM patterns. In S. I. Amari & M. Arbib (Eds.), *Competition and cooperation in neural networks.* New York: Springer-Verlag.

Grossberg, S. (1982b). The processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. *Psychological Review, 89,* 529–572.

Grossberg, S. (1982c). A psychophysiological theory of reinforcement, drive, motivation, and attention. *Journal of Theoretical Neurobiology, 1,* 286–369.

Grossberg, S. (1982d). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control.* Amsterdam: Reidel Press.

Grossberg, S. (1983). The quantized geometry of visual space: The coherent computation of depth form, and lightness. *Behavioral and Brain Sciences, 6,* 625–692.

Grossberg, S. (1984a). Some psychophysiological and pharmacological correlates of a developmental, cognitive, and motivational theory. In R. Karrer, J. Cohen, & P. Tueting (Eds.), *Brain and information: Event related potentials.* New York: New York Academy of Sciences.

Grossberg, S. (1984b). Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory, 7,* 263–283.

Grossberg, S. (1985). On the coordinated learning of item, order, and rhythm. Unpublished manuscript.

Grossberg, S., & Kuperstein, M. (1986). *Neural dynamics of adaptive sensory-motor control: Ballistic eye movements.* Amsterdam: North-Holland.

Grossberg, S., & Levine, D. S. (1975). Some developmental and attentional biases in the contrast enhancement and short term memory of recurrent neural networks. *Journal of Theoretical Biology, 53,* 341–380.

Grossberg, S., & Pepe, J. (1970). Schizophrenia: Possible dependence of associational span, bowing, and primacy vs. recency on spiking threshold. *Behavioral Science, 15,* 359–362.

Grossberg, S., & Pepe, J. (1971). Spiking threshold and overarousal effects in serial learning. *Journal of Statistical Physics, 3,* 95–125.

Grossberg, S., & Stone, G. O. (1986a). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review, 93,* 46–74.

Grossberg, S., & Stone, G. O. (1986b). Neural dynamics of attention switching and temporal order information in short term memory. Submitted for publication.

Halle, M., & Stevens, K.N. (1962). Speech recognition: A model and a program for research. *IRE Transactions and Information Theory, IT-8,* 155–159.

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

Haymaker, W., Anderson, E., & Nauta, W.J.H. (1969). *The hypothalamus,* Springfield, IL: C. C. Thomas.

Helson, H. (1964). *Adaptation level theory.* New York: Harper & Row.

Hoyle, G. (1977). *Identified neurons and behavior of arthropods.* New York: Plenum Press.

Johnston, J. C., & McClelland, J. L. (1974). Perception of letters in words: Seek not and ye shall find. *Science, 184,* 1192–1194.

Jusczyk, P. W. (1981). Infant speech perception: A critical appraisal. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology, 9,* 497–509.

Karrer, R., Cohen, J., & Tueting, P. (Eds.) (1984). *Brain and information: Event related potentials.* New York: New York Academy of Sciences.

Kelso, J.A.S., Southard, D. L., & Goodman, D. (1979). On the nature of human interlimb coordination. *Science, 203,* 1029–1031.

Kennedy, D. (1968). Input and output connection of single arthropod neurons. In F. O. Carlson (Ed.), *Physiological and biochemical aspects of nervous integration.* Englewood Cliffs, NJ: Prentice-Hall.

Kimura, D. (1976). The neural basis of language qua gesture. In H. Whitaker & H. A. Whitaker (Eds.), *Studies in neurolinguistics* (Vol. III). New York: Academic Press.

Kinsbourne, M., & Hicks, R. E. (1978). Mapping cerebral functional space: Competition and collaboration in human performance. In M. Kinsbourne (Ed.), *Asymmetrical function of the brain.* London: Cambridge University Press.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

Kuffler, S. W., & Nicholls, J. G. (1976). *From neuron to brain.* Sunderland, MA: Sinauer.

Lanze, M., Weisstein, N., & Harris, J. R. (1982). Perceived depth vs. structural relevance in the object-superiority effect. *Perception & Psychophysics, 31,* 376–382.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior.* New York: John Wiley.

Lawry, J. A., & LaBerge, D. (1981). Letter and word code interactions elicited by normally displayed words. *Perception & Psychophysics, 30,* 70–82.

Lenneberg, E. H. (1967). *Biological foundations of language.* New York: John Wiley.

Levine, D. S., & Grossberg, S. (1976). Visual illusions in neural networks: Line neutralization, tilt aftereffect, and angle expansion. *Journal of Theoretical Biology, 61,* 477–504.

Levinson, S. E., & Liberman, M. Y. (1981). Speech recognition by computer. *Scientific America,* April, 64–76.

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

Liberman, A. M., & Studdert-Kennedy, M. (1978). Phonetic perception. In R. Held, H. Leibowitz, & H. L. Teuber (Eds.), *Handbook of sensory physiology* (Vol. VIII). Heidelberg: Springer-Verlag.

Lindblom. B.E.F. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35,* 1773–1781.

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off skilled behavior. *Psychological Review, 89,* 483–506.

MacLean, P. D. (1970). The limbic brain in relation to psychoses. In P. Black (Ed.), *Physiological correlates of emotion*. New York: Academic Press.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69,* 548–558.

Marler, P. A. (1970). A comparative approach to vocal learning: song development in white-crowned sparrows. *Journal of Comparative and Physiological Psychology, 71,* 1–25.

Marler, P., & Peters, S. (1981). Birdsong and speech: Evidence for special processing. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189,* 226–228.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Marvilya, M. P. (1972). Spontaneous vocalizations and babbling in hearing-impaired infants. In C.G.M. Fant (Ed.), *Speech communication ability and profound deafness*. Washington, DC: A. G. Bell Association for the Deaf.

Matthei, E. H. (1983). Length effects in word perception: Comment on Samuel, van Santen, and Johnston. *Journal of Experimental Psychology, 9,* 318–320.

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

Miller, G. A. (1956). The magic number seven plus or minus two. *Psychological Review, 63,* 81–97.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions, In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25,* 457–465.

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

Murdock. B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Erlbaum.

Murdock, B. B. (1979). Convolution and correlation in perception and memory. In L. G. Nilsson (Ed.), *Perspectives in memory research: Essays in honor of Uppsala University's 500th anniversary*. Hillsdale, NJ: Erlbaum.

Myers, J. L., & Lorch, R. F. Jr. (1980). Interference and facilitation effects of primes upon verification processes. *Memory & Cognition, 8,* 405–414.

Neimark, E. D., & Estes, W. K. (Eds.). (1967). *Stimulus sampling theory*. San Francisco: Holden-Day.

Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman Press.

Newell, A. (1980). Harpy, production systems, and human cognition. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Norman, D. A. (1982). Categorization of action slips. *Psychological Review, 88,* 11–15.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44–64.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Olds, J. (1977). *Drives and reinforcements: Behavioral studies of hypothalamic functions*. New York: Raven Press.

Osgood, C. E. (1953). *Method and theory in experimental psychology*. New York: Oxford University Press.

Pastore, R. E. (1981). Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Patterson, P. H., & Purves, D. (Ed.). (1982). *Readings in developmental neurobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Lab.

Piaget, J. (1963). *The origins of intelligence in children*. New York: Norton.

Posner, M. I., & Snyder, C.R.R. (1975). Facilitation and inhibition in the processing of signals. In P.M.S. Rabbitt & S. Dornic (Eds.), *Attention and performance* (Vol. 5). New York: Academic Press.

Raaijmakers, J.G.W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93–134.

Ratcliff, R., & McKoon, G. (1981). Does activation really spread? *Psychological Review, 88*, 454–462.

Reeves, A., & Sperling, G. (1986). Attentional theory of order information in short-term visual memory. Preprint.

Repp, B. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language & Speech, 22*, 173–189.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of temporal cues for stop, fricative, and affricative manner. *Journal of Experimental Psychology, 4*, 621–637.

Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America, 69*, 1154–1163.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts.

Restle, F. (1978). Assimilation predicted by adaptation-level theory with variable weights. In N. J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3). Hillsdale, NJ: Erlbaum.

Robson, J. G. (1975). Receptive fields: Neural representation of the spatial and intensive attributes of the visual image. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. V). New York: Academic Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception, II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*, 60–94.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science, 6*, 1–36.

Samuel, A. G. van Santen, J.P.H., & Johnston, J. C. (1982). Length effects in word perception: We is better than I but worse than you or them. *Journal of Experimental Psychology, 8*, 91–105.

Samuel, A. G., van Santen, J.P.H., & Johnston, J. C. (1983). Reply to Matthei: We really is worse than you or them, and so are ma and pa. *Journal of Experimental Psychology, 9*, 321–322.

Sawusch, J. R., & Nusbaum, H. C. (1979). Contextual effects in vowel perception, I: Anchor-induced contrast effects. *Perception & Psychophysics, 25*, 292–302.

Sawusch, J. R., Nusbaum, H. C., & Schwab, E. C. (1980). Contextual effects in vowel perception II: Evidence for two processing mechanisms. *Perception & Psychophysics, 27*, 421–434.

Schneider, W., & Shiffrin, R. M. (1976). Automatic and controlled information processing in vision. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, NJ: Erlbaum.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic information processing I: Detection, search, and attention. *Psychological Review, 84*, 1–66.

Schwab, E. C., Sawusch, J. R., & Nusbaum, H. C. (1981). The role of second formant transitions in the stop-semivowel distinction. *Perception & Psychophysics, 29*, 121–128.

Semmes, J. (1968). Hemispheric specialization: A possible clue to mechanism. *Neuropychologia, 6*, 11-26.

Shaffer, L. H. (1982). Rhythm and timing in skill. *Psycghological Review, 89*, 109–122.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering, *Science, 210*, 390–398.

Smale, S. (1976). On the differential equations of species in competition. *Journal of Theoretical Biology, 3*, 5–7.

Soechting, J. F., & Laquaniti, F. (1981). Invariant characteristics of a pointing movement in man. *Journal of Neuroscience, 1*, 710–720.

Sperling, G., & Reeves, A. (1980). Measuring the reaction time of a shift of visual attention. In R. Nickerson (Ed.), *Attention and performance* (Vol. 7). Hillsdale, NJ: Erlbaum.

Squire, L. R., Cohen, N. J., & Nadel, L. (1982). The medial temporal region and memory consolidation: A new hypothesis. In H. Weingartner & E. Parker (Eds.), *Memory consolidation*. Hillsdale, NJ: Erlbaum.

Stein, L. (1958). Secondary reinforcement established with subcortical stimulation. *Science, 127*, 466–467.

Stein, P.S.G. (1971). Intersegmental coordination of swimmeret motoneuron activity in crayfish. *Journal of Neurophysiology, 34*, 310–318.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparison of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning*. New York: Academic Press.

Sternberg, S., Wright, C. E., Knoll, R. L., & Monsell, S. (1980). Motor programs in rapid speech: Additional evidence. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Stevens, C. F. (1966). *Neurophysiology: A primer*. New York: John Wiley.

Stevens, K. N. (1972). Segments, features, and analysis by synthesis. In J. V. Cavanaugh & I. G. Mattingly (Eds.), *Language by eye and by ear*. Cambridge, MA: MIT Press.

Stevens, K. N., & Halle, M. (1964). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Proceedings of the AFCRL symposium on models for the perception of speech and visual form*. Cambridge, MA MIT Press.

Studdert-Kennedy, M. (1975). The nature and function of phonetic categories. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, NJ: Erlbaum.

Studdert-Kennedy, M. (1980). Speech perception. *Language & Speech, 23*, 45–65.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77*, 234–249.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135–170.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 393–395.

Warren, R. M., & Obusek, D. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics, 9,* 358–362.

Watkins, O. C., & Watkins, M. J. (1982). Lateral inhibition and echoic memory: Some comments on Crowder's (1978) theory. *Memory & Cognition, 10,* 279–286.

Weisstein, N. (1968). A Rashevsky-Landahl neural net: Simulation of metacontrast. *Psychological Review, 75,* 494–521.

Weisstein, N. (1972). Metacontrast. In D. Jameson & L. M. Hurvich (Eds.), *Handbook of sensory physiology* (Vol. VII/4). Berlin: Springer-Verlag.

Welford, A. T. (1968). *Fundamentals of skill.* London: Methuen.

West, M. O., Christian, E., Robinson, J. H., & Deadwyler, S. A. (1981). Dentate granule cell discharge during conditioning. *Experimental Brain Research, 44,* 287–294.

Willows, A. O. D. (1968). Behavioral acts elicited by stimulation of single identifiable nerve cells. In F. O. Carlson (Ed.), *Physiological and biochemical aspects of nervous integration.* Englewood Cliffs, NJ: Prentice-Hall.