In Evolution, Learning, Cognition, and Advanced Architectures, Y.C. Lee (Ed.). Hong Kong: World Scientific Publishers, 1988.

SPEECH PERCEPTION AND PRODUCTION BY A SELF-ORGANIZING NEURAL NETWORK

Michael A. Cohen*, Stephen Grossberg*, and David G. Stork*+

Ctr. Adaptive Systems (*)
Boston University
111 Cummington Street
Boston, MA 02215

Department of Physics (+)
and Program in Neuroscience
Clark University
Worcester, MA 01610

Abstract

Considerations of the real-time self-organization of neural networks for speech recognition and production have lead to a new understanding of several key issues in such networks, most notably a definition of new processing units and functions of hierarchical levels in the auditory system. An important function of a particular neural level in the auditory system is to provide a partially-compressed code, mapped to the articulatory system, to permit imitation of novel sounds. Furthermore, top-down priming signals from the articulatory system to the auditory system help to stabilize the emerging auditory code. These structures help explain results from the motor theory, which states that speech is analyzed by how it would be produced. Higher stages of processing require chunking or unitization of the emerging language code, an example of a classical grouping problem. The partially compressed auditory codes are further compressed into item codes (e.g., phonemic segments), which are stored in a working memory representation whose short-term memory pattern is its code. A masking field level receives input from this working memory and encodes this input into list chunks, whose top-down signals organize the items in working memory into coherent groupings with invariant properties. This total architecture sheds new light on key speech issues such as coarticulation, analysis-by-synthesis, motor theory, categorical perception, invariant speech perception, word superiority, and phonemic restoration.

1. The Learning of Language Units

During a human's early years, an exquisitely subtle and sensitive speech recognition and production system develops. These two systems develop to be well-matched to each other, enabling rapid and reliable broadcast and reception of linguistic information. The development of these systems can be viewed as resulting from two fundamental processes: self-organization through *circular reaction* and through *chunking* or *unitization*. This chapter sketches some issues concerning these processes in speech and provides a summary of its key neural components, developed to address more general cognitive problems.

2. Low Stages of Processing: Circular Reactions and the Emerging Auditory and Motor Codes

The concept of circular reaction (Piaget, 1963) is illustrated in Figure 1. For our purposes, the reaction links the *motor* or *articulatory* system (mouth, tongue, velum, etc., and the neural structures controlling them) with the *auditory* system (ear and its neural

M.A.C. was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-86-C-0037) and the National Science Foundation (NSF IRI-84-17756), S.G. was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-86-C-0037 and AFOSR F49620-87-C-0018), and D.G.S. was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-86-C-0037).

Acknowledgements: We wish to thank Cynthia Suchta and Carol Yanakakis for their valuable assistance in the preparation of the manuscript.

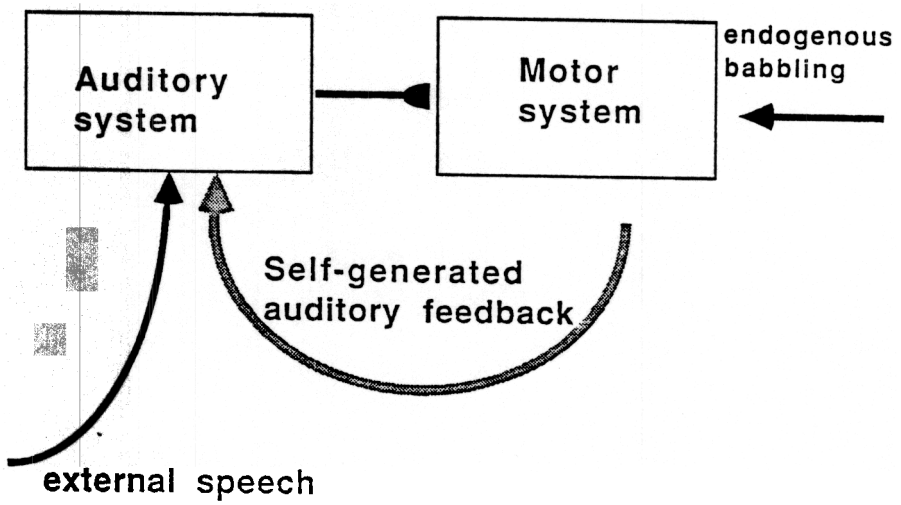

Figure 1. Circular reaction linking the motor system to the auditory system. Such a loop permits imitation of novel sounds from an external speaker.

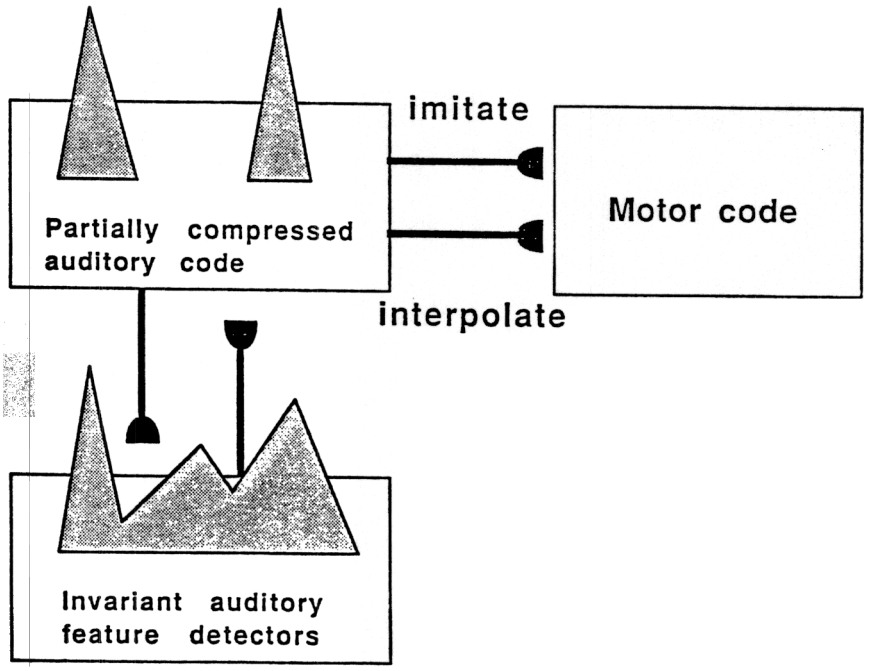

Figure 2. Neural interactions between a partially-compressed auditory code and a motor code permits the imitation of novel heard sounds.

perceptual mechanisms). In a developing infant, endogenously generated babbling signals in the motor system lead to auditory feedback, thereby allowing the auditory system to tune its evolving recognition codes. Moreover, the auditory system can compare the self-generated sounds to those from external speakers.

Figure 2 shows in slightly greater detail relevant neural interconnections in the auditory model. After processing by low-level auditory feature detectors (detecting energy in various frequency wavebands, "sweeping" frequency signals, broad-band or burst energy distributions) the auditory information is partially compressed and passed to a subsequent level, where it is represented by significant activity in a smaller number of neurons.

There is a learned auditory-to-articulatory associative map at this level, important for the following purposes. First, it permits the motor system to *interpolate* novel heard sounds. That is, if a novel sound leads to an auditory code "between" those for other, previously coded sounds, then this novel sound will be mapped to a motor code "between" those for the sounds previously heard. Second, the associative map permits the motor system to *imitate* such sounds. In this manner, a novel sound will lead to a novel, interpolated motor code. When accessed, this new motor code will lead to an utterance closer to the novel one heard. This (imitated) utterance then accesses an auditory code very similar to the interpolated one.

The auditory code at the level for this interpolation and imitation must be only *partially* compressed; a fully compressed (or *unitized*) code would map to a previously organized motor code, precluding interpolation of novel sounds. Furthermore, the auditory level for interpolation must be above stages of invariant preprocessing—only in this way can effects such as vocal tract normalization be explained (Lieberman, 1984, pp.219–223). It has been argued (Lieberman, 1984, p.222) that such normalization is due to the existence of innate mechanisms, and hence is not modifiable in the manner of the auditory-to-motor map.

3. The Vector Integration to Endpoint Model

The motor code in our network is based on the recent Vector Integration To Endpoint (VITE) model of arm movement control (Bullock and Grossberg, 1987), due to functional similarities between speech articulation and arm movement problems. Moreover, we agree with Lieberman (1984) that phylogenetically the speech system appropriated the speech articulators and their neural controlling structures from their original tasks of swallowing, chewing, and so forth—tasks more typical of standard motor control concerns. The VITE model posits three interacting neural levels: (1) a Target Position Command (TPC) level, whose spatial distribution of activity codes where the limb "wants to go," (2) a Present Position Command (PPC) level, which generates an outflow movement command, and (3) a Difference Vector (DV) level, which compares the TPC and PPC codes. Such a structure has been used to explain a range of motor control psychophysics and physiology results, in particular (for our speech system) the simultaneous contraction of several muscle groups in a synergy, even at different overall rates. The learning of a motor task, in this scheme, involves the printing (i.e., modification of synapses for long-term memory) of the motor code when the limb is at or near the target position. Put another way, learning occurs when the present position and the target position form a near match (i.e., when $DV < \epsilon$). Hence in our speech system the Difference Vector layer can act as a learning *gate*, regulating the formation of the auditory-to-articulatory map during the near match condition, as shown in Figure 3.

Speech articulators, however, do not all function as a single, unitized system; rather, there are several muscle synergies or *coordinative structures* (Fowler, 1980) working quasi-independently. For instance, one coordinative structure might link the jaw and front of the tongue for bringing the top of the tongue to the hard palate in order to utter [t], while a different coordinative structure is controlling the back of the tongue to utter a (coarticulated) [a]. Each of the coordinative structures must have its own TPC, PPC,

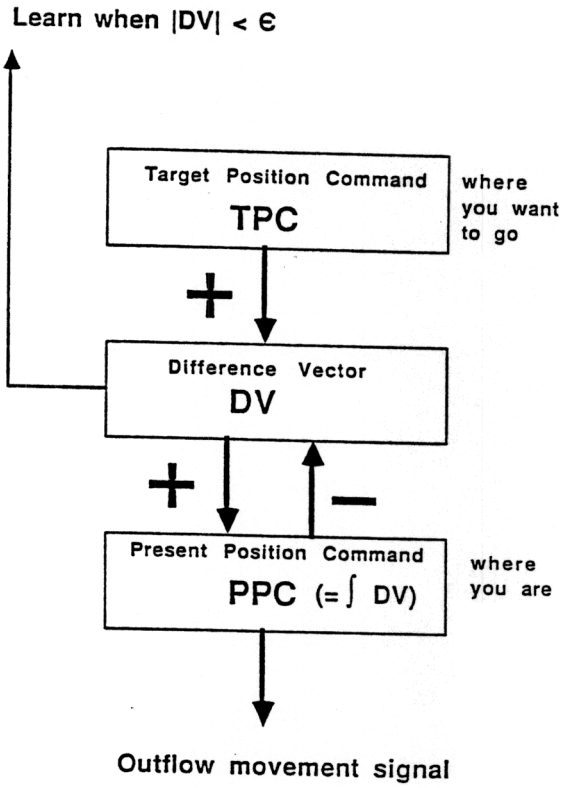

Figure 3. Basic VITE module and its learning gate, for use in encoding the TPC codes.

and DV layers, to preserve such quasi-independence. Figure 4 shows how the TPC's of different coordinative structures are chunked into distinct motor control commands. Thus the imitative map can associate different aspects of the partially compressed auditory code with different coordinative structures. Figure 4 also shows the basic structure of the circular reaction loop linking the auditory system and the motor system, incorporating the VITE circuit and its learning gate.

4. Self-Stabilization of Imitation via Motor-to-Auditory Priming

In a self-organized system, a key issue concerns the ability of the system to *self-stabilize* its learning under natural conditions (Carpenter and Grossberg, 1987a, 1987b). During speech the auditory code varies (in general) *continuously* due to its representation of a stream of varying sounds, whereas the controlling motor code varies more *discretely* due to the fact that new target position commands (TPCs) are printed by the imitative associative map only when the motor system achieves an approximate match (Figure 3), either at an initial TPC or a final TPC of a simple utterance (Figure 5). This raises the issue of insuring that the emerging auditory code is *consistent* with the motor code so that the imitative map can self-stabilize. Such consistency can be achieved through top-down motor priming which associates the compressed motor codes that represent the coordinative structures with activation patterns across the auditory feature detectors, as shown in Figure 6—an example of active internal regulation by top-down resonant feedback.

The top-down motor expectations (or priming signals) reorganize the auditory code to make it consistent with the evolving motor code. Such priming occurs during the activity of any given motor code, and hence reinforces the activity patterns across auditory feature detectors that are heard contemporaneous and consistent with such motor codes. These motorically-modified feature activity patterns are encoded in long-term memory within

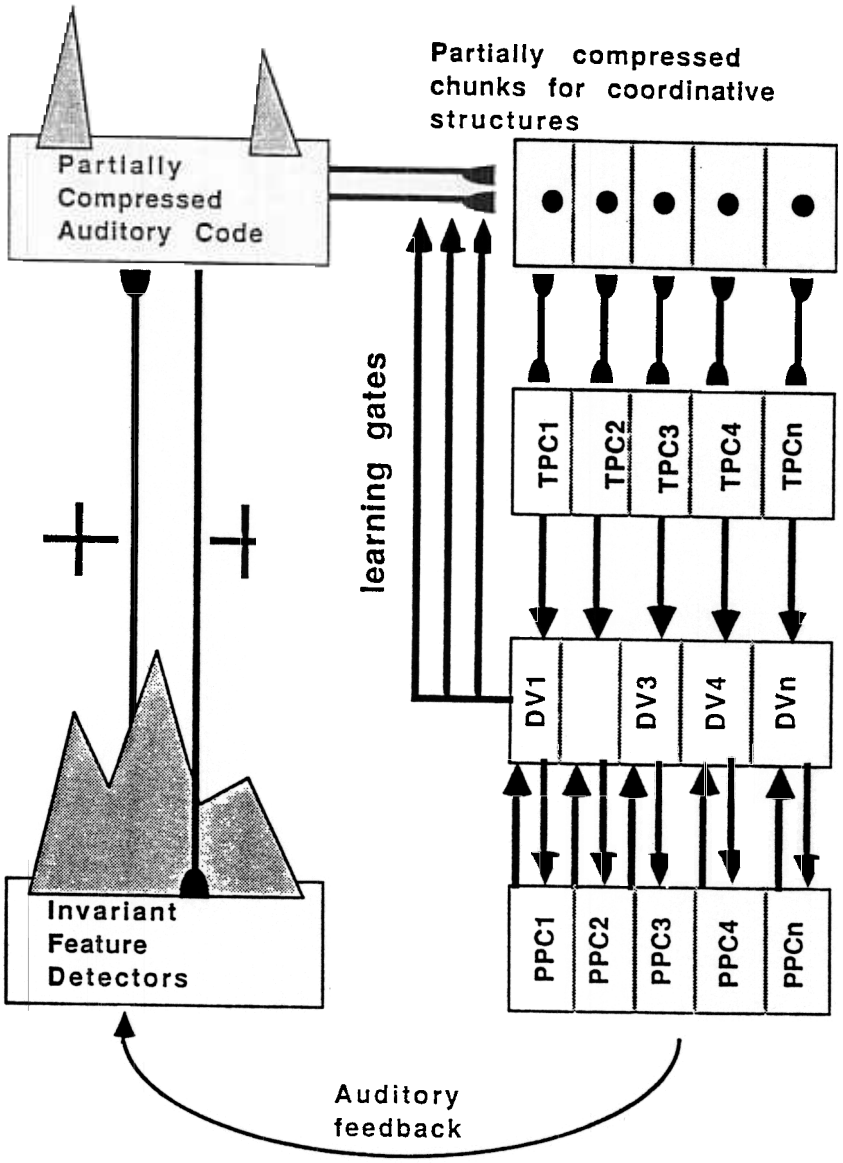

Figure 4. Circular reaction loop linking the motor system (right) with auditory system (left). Parallel motor channels for coordinative structures are shown, each with its associated learning gate, which prints (modifies the synapses for long-term memory) the imitative map between the partially-compressed auditory code and the motor code.

the auditory-to-auditory pathways to the partially compressed auditory code. Even during passive listening, these motorically-influenced auditory codes are activated. Heard speech is thereby analyzed by "how it would have been phonated." This is in agreement with the motor theory of speech perception (c.f., Studdert-Kennedy, 1984) and finds support from physiology (Ojemann, 1983). These results and the architecture of Figure 6 clarify why the concerted attempts to find purely auditory correlates of speech segments have not met with greater success (c.f., Zue, 1976; Cooper, 1980, 1983), and suggests how an artificial system capable of recognizing natural speech can incorporate motor information that human listeners employ.

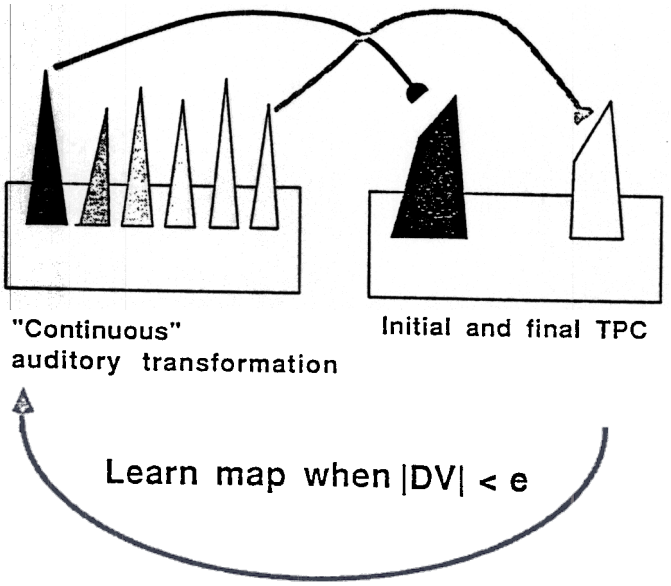

Figure 5. The dynamic pattern of the code in the auditory system is more continuous, while that in the control structure for the motor system is more discrete. When activated, such a motor code initiates a unitized, stereotyped synergetic action of articulators.

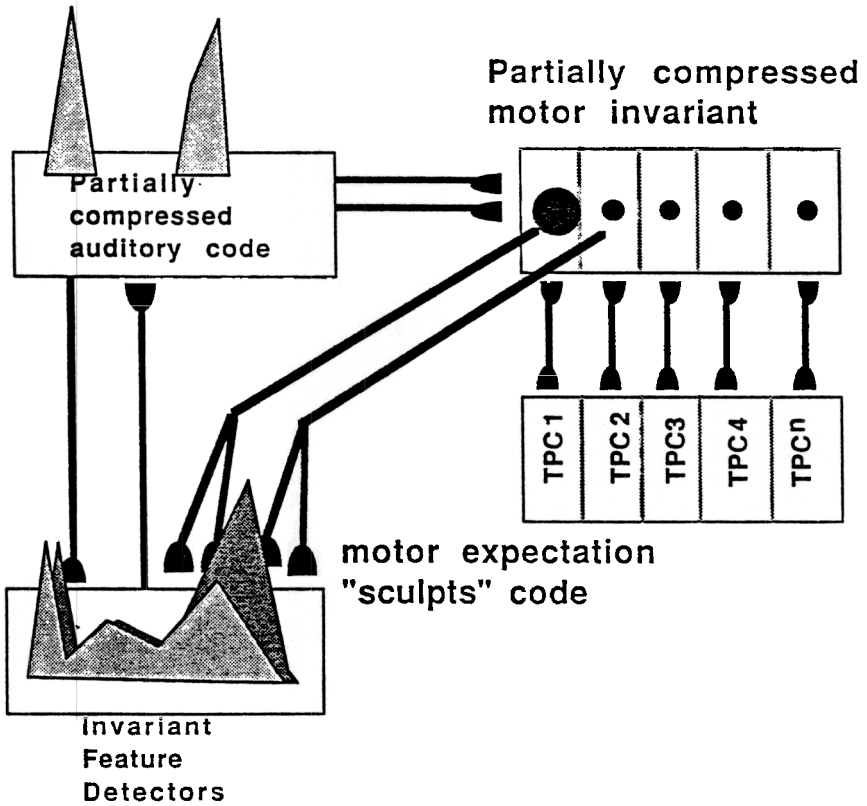

Figure 6. Top-down priming from the motor to the auditory system reorganizes the emerging auditory code to be consistent with the motor commands. This motorically-influenced auditory code is further compressed at higher stages of the auditory system.

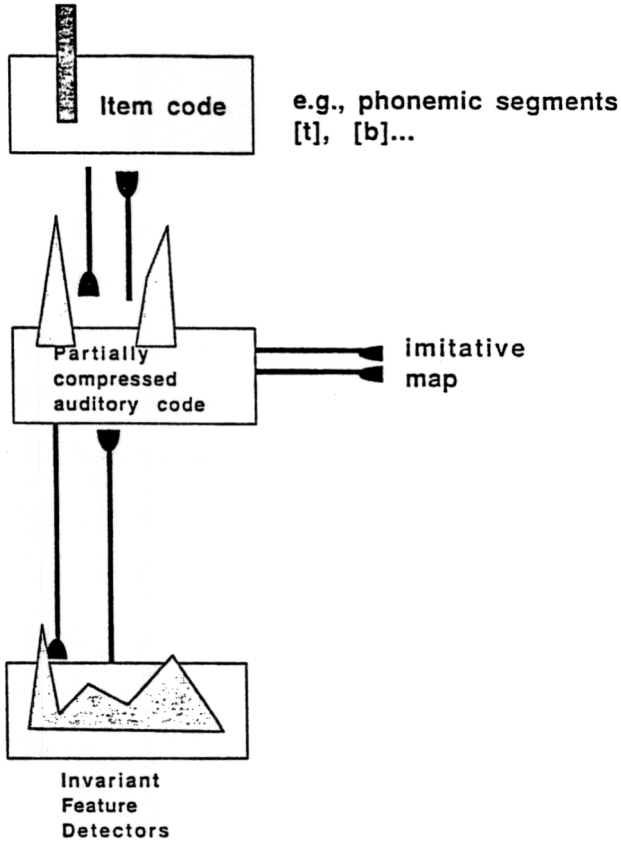

Figure 7. Unitization is achieved by compressing the partially compressed auditory code to yield an item code, which includes such units as phonemic segments.

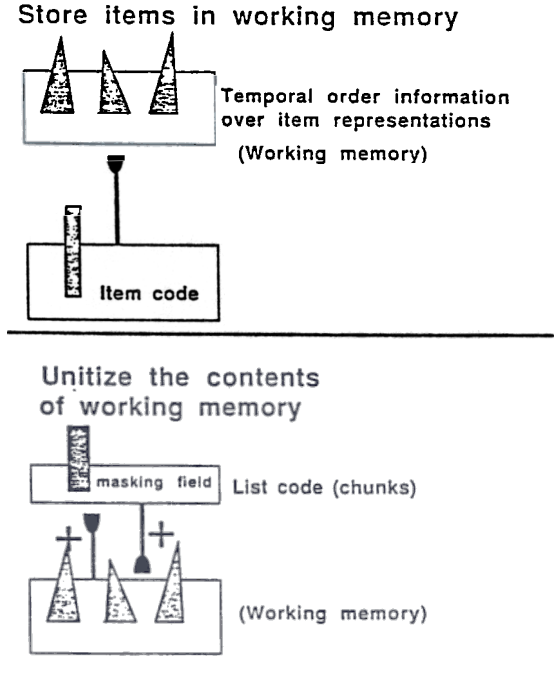

Figure 8. Context-sensitive list codes are formed via a two-level process: (top) Items are placed in *working memory*, which encodes temporal order information. Then (bottom) a masking field uses bottom-up flow and top-down priming to yield context-sensitive list codes.

5. Higher Stages of Processing: Context-Sensitive Chunking and Unitization of the Emerging Auditory Speech Code

Stages of the auditory system higher than the ones described above rely on processes other than circular reactions for stabilizing the emerging language code. Such processes *unitize*, *chunk*, or *group* the emerging discrete linguistic units in a context-sensitive manner. Such context-sensitivity is crucial if the network is to be able to classify any given phonemic segment (say) in all its coarticulated forms.

An early stage of unitization is achieved by compressing the partially compressed auditory code to yield an *item code*, as shown in Figure 7. Grouping such items into context-sensitive chunks requires two stages, as shown in Figure 8. First, sequentially occurring items are stored in a *working memory* level to encode temporal order information over the items. Next, these items are grouped by a *masking field* (Cohen and Grossberg, 1986, 1987) into context-sensitive list chunks.

6. Masking Fields

In brief, a masking field neural structure possesses both bottom-up and top-down interconnections with the item level (Figure 9). Nodes at the list level compete through mutual inhibition. List nodes that are best predictive of *longer* patterns of items will inhibit the less predictive nodes for shorter lists. Recognition of a unitized grouping of items occurs when a bottom-up top-down context-sensitive *resonance* develops. In speech networks, such a masking field can thus unitize the evolving auditory code into predictive chunks, representing, say, phonemic segments.

Figure 10 schematizes the anatomy of a masking field. Figure 11 schematizes the two primary types of coding sensitivity of which a masking field is capable in response to bottom-up inputs from an item field. Figures 12 and 13 summarize computer simulations which demonstrate this coding competence.

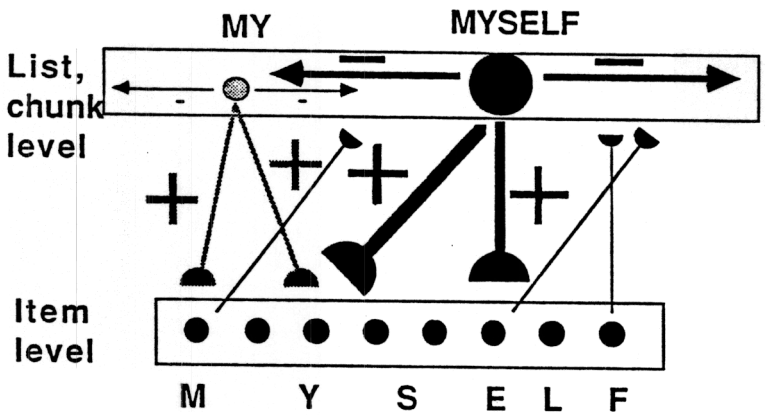

Figure 9. A masking field architecture creates context-sensitive list codes by using both bottom-up filtering signals and top-down priming signals from the list level. There is competition between units in the list level. "Larger" nodes—ones that pool information from a larger number of items—inhibit "smaller" nodes more effectively than vice versa. For instance, if list nodes for MY, SELF, ELF, and MYSELF are encoded, the presentation of the letters M-Y-S-E-L-F at the item level will lead to a resonance between the MYSELF node and the six items, while nodes representing smaller, less predictive, groupings are quickly suppressed.

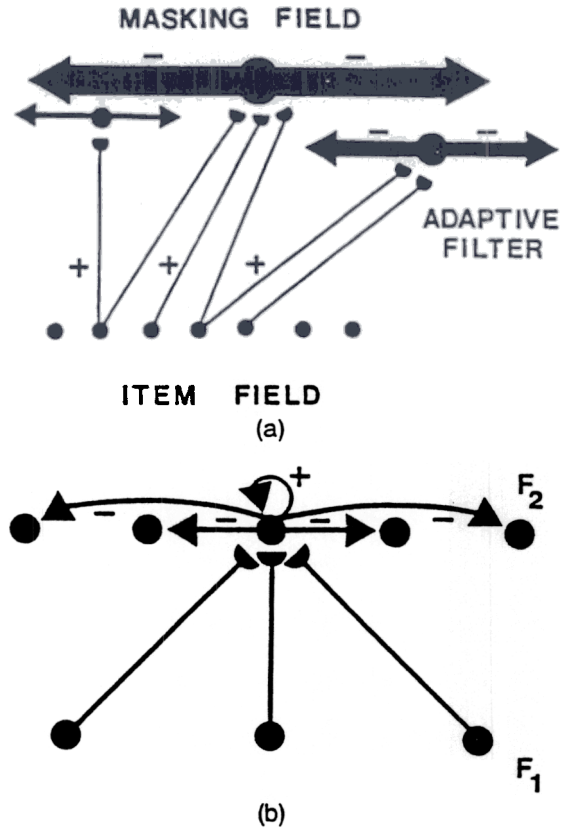

Figure 10. Masking field interactions: (a) Cells from an item field F_1 grow randomly to a masking field F_2 along positionally sensitive gradients. The nodes in the masking field grow so that larger item groupings, up to some optimal size, can activate nodes with broader and stronger inhibitory interactions. Thus the $F_1 \rightarrow F_2$ connections and the $F_2 \leftrightarrow F_2$ interactions exhibit properties of self-similarity. (b) The interactions within a masking field F_2 include positive feedback from a node to itself and negative feedback from a node to its neighbors. Long term memory (LTM) traces at the ends of $F_1 \rightarrow F_2$ pathways (designated by hemidisks) adaptively tune the filter defined by these pathways to amplify the F_2 reaction to item groupings which have previously succeeded in activating their target F_2 nodes.

The interactions between these levels can explain many speech properties, including properties of temporal invariance and phonemic restoration. When designed to incorporate a "long-term memory invariance principle" (Grossberg, 1986, 1987; Grossberg and Stone, 1986a, 1986b), the spatial pattern of activation across working memory defines an invariant code, and an attentional gain control signal to the working memory stage preserves this spatial code under changes in overall speaking rate.

Phonemic restoration occurs when an ambiguous or missing sound is clearly heard when presented in the proper context. The top-down priming of a masking field can complete ambiguous elements of the item code, so long as these items can be reorganized by the 2/3 Rule properties of the prime (Carpenter and Grossberg, 1987a, 1987b). The speech code results from a resonant wave which is controlled by feedback interactions between the working memory and masking field levels. Although the list chunks which reorganize the form and grouping of item codes utilize "future" information, this resonant wave can emerge from "past" to "future" because the internal masking of unpredictive list codes within the masking field occurs much faster than the time scale for unfolding the resonant

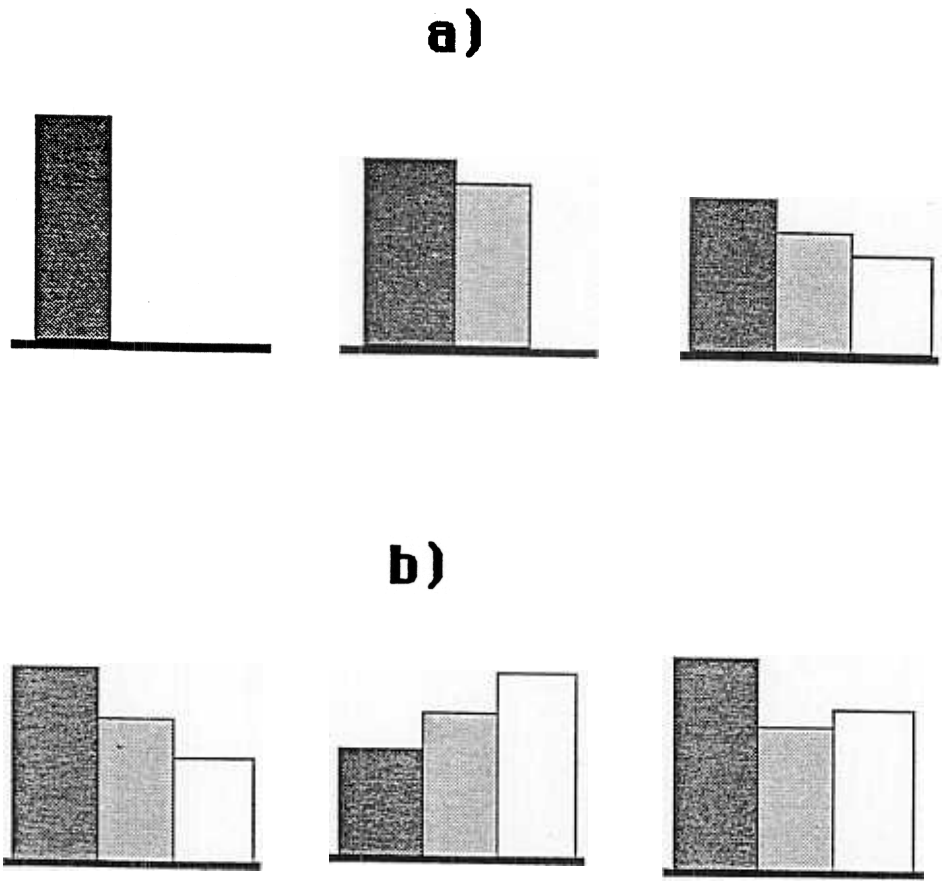

Figure 11. Two types of masking field sensitivity: (a) A masking field F_2 can automatically rescale its sensitivity to differentially react as the F_1 activity pattern expands through time to activate more F_1 cells. It hereby acts like a "multiple spatial frequency filter." (b) A masking field can differentially react to different F_1 activity patterns which activate the same set of F_1 cells. By (a) and (b), F_2 acts like a spatial pattern discriminator which can compensate for changes in overall spatial scale without losing its sensitivity to pattern changes at the finest spatial scale.

wave (Figure 14.)

The overall neural architecture employing the elements described above is shown in Figure 15.

Additional network designs are being developed for dealing with additional problems such as factoring rhythm information from linguistic information and the coding of repetitive patterns. Even as it stands, however, the architecture and design considerations described above provide a new processing architecture for understanding such issues as analysis-by-synthesis, the motor theory of speech perception, categorical perception, invariant speech perception, and phonemic restoration.

References

Bullock, D. and Grossberg, S., Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, in press, 1987.

Carpenter, G.A. and Grossberg, S., A massively parallel architecture for a self-organizing

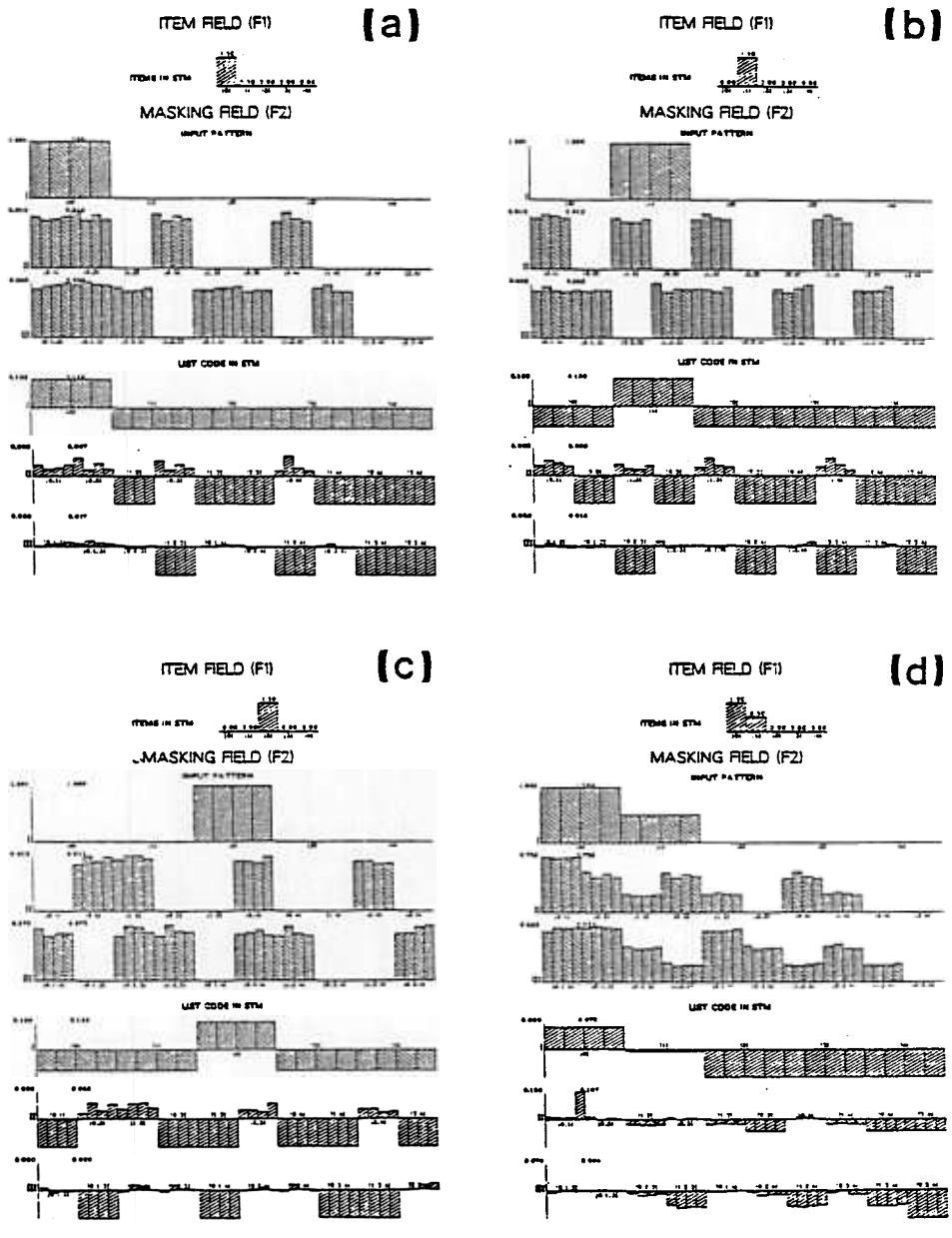

Figure 12. (a) The correct list code $\{0\}$ is preferred in STM, but predictive list codes which include $\{0\}$ as a part are also activated with lesser STM weights. The prediction gets less activation if $\{0\}$ forms a smaller part of it. (b) The correct list code $\{1\}$ is preferred in STM, but the predictive list codes which include $\{1\}$ as a part are also activated with lesser STM weights. (c) The list code in response to item $\{2\}$ also generates an appropriate reaction. (d) A list code of type $\{0,1\}$ is maximally activated, but part codes $\{0\}$ and predictive codes which include $\{0,1\}$ as a part are also activated with lesser STM weights.

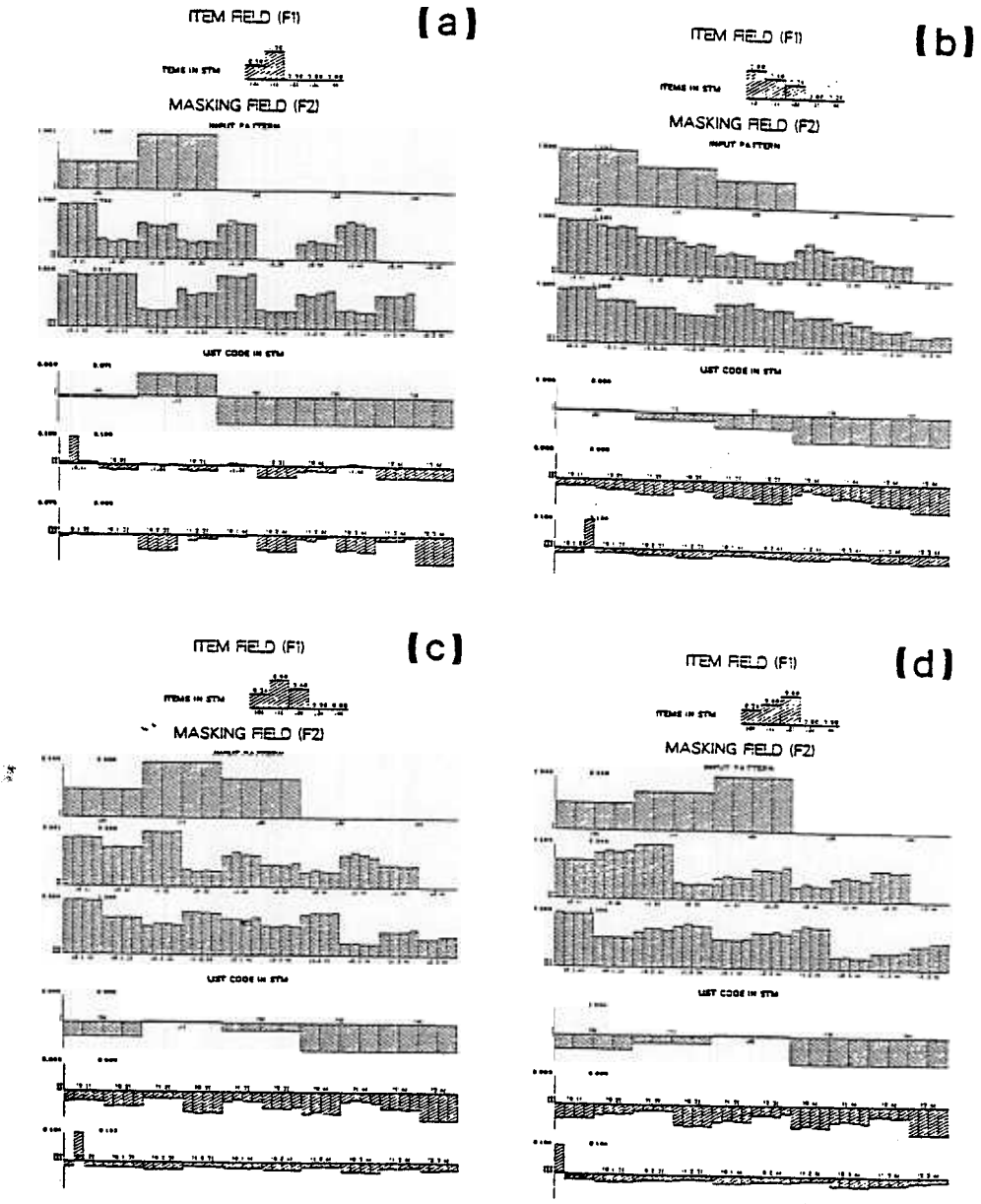

Figure 13. (a) A different list code of type $\{0,1\}$ is maximally activated, but part codes $\{1\}$ are also activated with lesser STM weight. Due to the random growth of $F_1 \rightarrow F_2$ pathways, no predictive list codes are activated (to 3 significant digits). (b)–(d) When the STM pattern across F_1 includes three items, the list code in STM strongly activates an appropriate list code. Part groupings are suppressed due to the high level of predictiveness of this list code. Comparison of Figures 12a, 12d, and 13b shows that as the item code across F_1 becomes more constraining, the list code representation becomes less distributed across F_2.

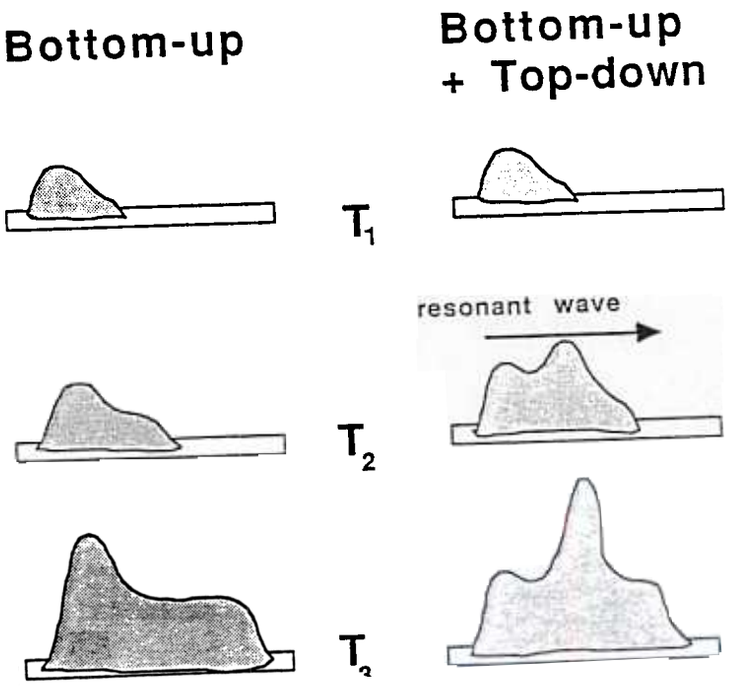

Figure 14. (Left): The activity pattern in working memory as new items enter the system, if the architecture had purely bottom-up connections. (Right): If the system has top-down priming, on the other hand, crucial features in the working memory that fit into a coherent pattern are reinforced, leading to a different distribution of neural activity. This resonant wave constitutes the speech code.

neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 1987, **37**, 54–115 (a).

Carpenter, G.A. and Grossberg, S., ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, in press, 1987 (b).

Cohen, M.A. and Grossberg, S., Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, 1986, **5**, 1–22.

Cohen, M.A. and Grossberg, S., Masking fields: A massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, 1987, **26**, 1866–1891.

Cooper, F.S., Acoustics in human communication: Evolving ideas about the nature of speech. *Journal of the Acoustical Society of America*, 1980, **68**, 18–21.

Cooper, F.S., Some reflections on speech research. In P.F. MacNeilage (Ed.), **The production of speech**. New York: Springer-Verlag, 1983.

Fowler, C., Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 1980, **8**, 113–133.

Grossberg, S., The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Pattern recognition by**

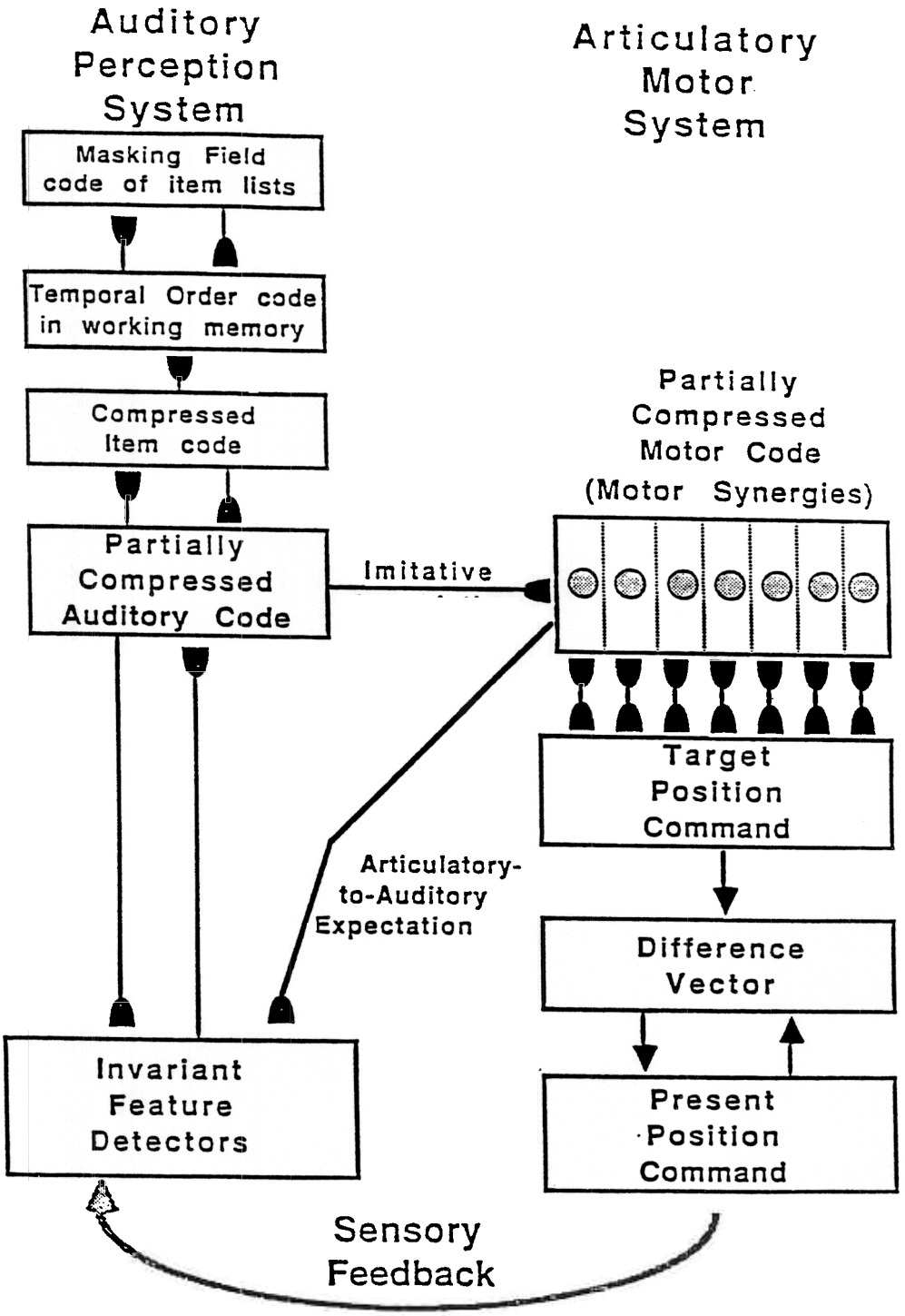

Figure 15. Global architecture for a speech recognition and synthesis system, employing the processing described above. See text for details.

humans and machines, Vol. 1: Speech perception. New York: Academic Press, 1986.

Grossberg, S. (Ed.), **The adaptive brain, II: Vision, speech, language, and motor control.** Amsterdam: Elsevier/North-Holland, 1987.

Grossberg, S. and Stone, G.O., Neural dynamics of attention switching and temporal order information in short-term memory. *Memory and Cognition*, 1986, **14**, 451–468 (a).

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

Lieberman, P., **The biology and evolution of language.** Cambridge, MA: Harvard University Press, 1984.

Ojemann, G., Brain organization for language from the perspective of electrical stimulation mapping. *Behavioral and Brain Sciences*, 1983, **2**, 189–230.

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

Studdert-Kennedy, M., Perceptual processing links to the motor system. In M. Studdert-Kennedy (Ed.), **Psychobiology of language.** Cambridge, MA: MIT Press, 1984, 29–39.

Zue, V.W., Acoustic characteristics of stop consonants: A controlled study. Ph.D. Dissertation, Massachusetts Institute of Technology, Electrical Engineering and Computer Science Department, 1976.