Self-Organization of Auditory Motion Detectors

Sven E. Anderson*
Department of Linguistics
Indiana University
Bloomington, Indiana 47405
sven@cs.indiana.edu

Abstract

This work addresses the question of how neural networks self-organise to recognise familiar sequential patterns. A neural network model with mild constraints on its initial architecture learns to encode the direction of spectral motion as auditory stimuli excite the units in a tonotopically arranged input layer like that found after peripheral processing by the cochlea. The network consists of a series of inhibitory clusters with excitatory interconnections that self-organize as streams of stimuli excite the clusters over time. Self-organization is achieved by application of the learning heuristics developed by Marshall (1990) for the self-organization of excitatory and inhibitory pathways in visual motion detection. These heuristics are implemented through linear thresholding equations for unit activation having faster-than-linear inhibitory response. Synaptic weights are learned throughout processing according to the competitive algorithm explored in Malsburg (1973).

The Perception of Spectral Motion

The processing of sequential stimuli is an essential component of auditory and visual perception in many animals. Recent efforts have resulted in learning algorithms that can be used to encode sequential patterns within autoassociative (Reiss & Taylor, 1991; Metzger & Lehmann, 1990; Elman, 1990) and supervised paradigms (Wang & Arbib, 1990; Földiák, 1991). We believe that these approaches can be successfully extended to the self-organization of sequential pattern detectors through the integration of a hierarchy of network layers, each of which is sensitive to particular attributes of a sequential input stream. This report details an implementation of the first module of a system for building representations of sequential auditory patterns that are statistically salient in an animal's environment. When exposed to an environment consisting of frequency sweeps, sound bursts, and constant frequency components this module learns to detect direction of motion from a 1-dimensional tonotopic input array.

Neural patterns of response to auditory stimuli travel from the basilar membrane to the auditory cortex via the cochlear nucleus, inferior colliculus, and medial geniculate. There is little doubt that higher and higher centers of auditory processing respond to auditory stimuli of increasing complexity and duration (Pickles, 1988). Stimuli with changing frequency are important to many species for communication, navigation, and target tracking. For example, within the mustached bat, sensitivity to frequency modulated tones (FM) has been found at the cochlear nucleus (Suga, 1990). Auditory cortex apparently contains large numbers of units that respond best to species specific calls (Aitkin, 1990) which are usually temporally and spectrally complex. Whitfield and Evans (1965) discovered that the majority of a sample of 104 cells of auditory cortex responded only to frequency modulation in a particular direction. The effect of rate of frequency modulation on cell response was minimal though perceptible for some cells. We thus propose that an important first task of the auditory pathway is the rate-invariant determination of direction of motion across the spectrum for non-constant stimuli.

The motion detection model presented here converts an inherently temporal pattern into a spatial code (unit activity). This may be useful if further processing is to isolate sequential patterns using spatial learning mechanisms like competitive learning or the delta rule. Since the non-stationary aspect of signals is more important to speech that steady frequency components, direct representation of frequency change emphasizes functionally relevant aspects of acoustic signals. Finally, the predictive aspect of motion detection should permit sequence tracking to be robust in the complex acoustic environment faced by most animals.

Auditory Motion Layer

Preprocessing

For testing on actual auditory stimuli, input to the model approximates response characteristics of the auditory nerve. These characteristics could be modeled using a model like that studied in

*Supported by the Indiana University Graduate School and the Armed Forces Communications and Electronics Association. This work was also supported by ONR grant N00014-91-J1261 to Robert Port.

(Delgutte, 1982), but are merely simulated here. The important property of the preprocessor is that it consists of an array of linear bandpass filters, each followed by adaptation that leads to rapid ON-type response and then decay to a much lower value. Consequently, the input stimuli used in these simulations consists of sequences of binary-valued patterns sweeping across the input field as though ON-type responses had been filtered through a cutoff-threshold.

Model and Learning

The topology and learning heuristics of the present model are adapted from those presented in (Marshall, 1990) for the processing of visual motion and velocity information. Marshall's model employed the shunting equations studied by Grossberg (1973) and the competitive learning equations outlined in (Carpenter & Grossberg, 1987). In early simulations it was found that the shunting equations proposed by Marshall contain a number of strong linearities that require careful numerical integration and are therefore computationally expensive. These equations were revised in order to make possible the eventual simulation of much larger networks necessary to speech processing over the entire audible spectrum. We found that the essential features of Marshall's model are retained in the current formulation.

The motion detection layer (Figure 1) is a tonotopic layer of inhibitory clusters, the units of which are connected to the units of all other inhibitory clusters in the same layer by excitatory connections having fixed delay. The clusters themselves are on-center off-surround anatomies that emphasize the activation of the unit with greatest activation. Each input line connects to all units in a single cluster corresponding to the receptive field represented by the input line, thus preserving the tonotopic arrangement of the input units.

Initially the lateral excitatory connections between units of the motion detection layer are randomly connection. Over time these connections organize themselves to represent the spatiotemporal correlations present in the input environment. Learning is proportional to the degree to which bottom-up input to a unit coincides with input from units in other clusters. A unit that receives both bottom-up and lateral excitation tends to suppress other units in its cluster and, as a result, learns more strongly than other units in its cluster. Competition between units ensures that the units of each cluster respond to different input patterns.

Unit Equations. The essential attributes of units in the inhibitory clusters are determined by the necessity that all units in a cluster activate in the presence of bottom-up input, and the opposing requirement that combined lateral and bottom-up excitation cause winner-take-all behavior. Moreover, the selection of the most strongly activated unit in a cluster must be rapid, or intermediate activation values will corrupt learning. One can distill

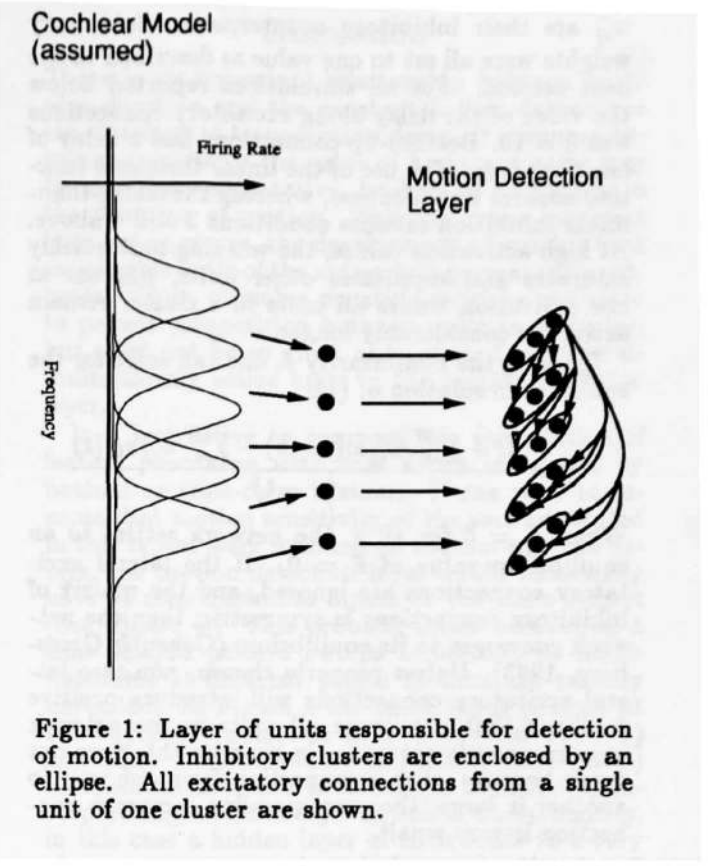

Figure 1: Layer of units responsible for detection of motion. Inhibitory clusters are enclosed by an ellipse. All excitatory connections from a single unit of one cluster are shown.

the essential behavior of Marshall's motion detection model to the following four points:

1. The network must be stable.
2. When input to a unit falls to zero the unit's activation must rapidly decay to zero.
3. At low activation values units in a cluster can be simultaneously active.
4. At high activations the unit having greatest activation rapidly saturates while simultaneously suppressing other units in a cluster.

All units in the motion detection layer obey the equation

$$(1)\ x_j(t+1) = x_j(t)[1-\gamma\tau\Delta t] + \tau\Delta t\, f(I_j + x_j(t) + \sum_i^N w_{ij}^+ x_i(t-k) - \sum_i^N w_{ij}^-(1+x_i(t))^2)$$

where f is the linear threshold function

$$f(z) = \begin{cases} 0 & z \le 0 \\ z & 0 < z < 1 \\ 1 & z \ge 1 \end{cases}$$

The unit activation function is the Euler approximation to the corresponding differential equation, and discretization is controlled by the value of Δt. The parameters τ and γ are the time constant of a unit and its decay rate, respectively. These parameters are the same for all units. The w_{ij}^+ are excitatory synaptic weights from unit i to j, and the

w_{ij}^- are their inhibitory counterparts. Inhibitory weights were all set to one value as described in the next section. For all simulations reported below the value of the delay along excitatory connections was $k = 10$. Bottom-up connections had a delay of one time step. The use of the linear threshold function ensures boundedness, whereas the faster-than-linear inhibition satisfies conditions 3 and 4 above. At high activation values, the winning unit quickly saturates and suppresses other units, whereas at low activation values all units in a cluster remain active for considerably longer.

Ignoring the nonlinearity f, one can solve for the equilibrium solution of (1).

$$x_j = \frac{I_j + \sum_i^N w_{ij}^+ x_i(t-k) - \sum_i^N w_{ij}^- x_i^2(t)}{(\gamma - 1)}$$

When $I_j = 0$ for all j, the network settles to an equilibrium value of $\vec{x} = 0$. If the lateral excitatory connections are ignored, and the matrix of inhibitory connections is symmetric, then the network converges to its equilibrium (Cohen & Grossberg, 1983). Unless properly chosen, non-zero lateral excitatory connections will introduce positive feedback that can cause all units in the network to permanently saturate. In practice this does not occur because when a connection from one unit to another is large, the corresponding recurrent connection is very small.

Network Initialization. Initially we set all of the inhibitory weights within clusters to $\frac{-1}{N_c}$, where N_c is the number of units per cluster. Excitatory weights between all units outside a cluster favor local connections and were set to

$$w_{ij}^+ = (1.0 + r)e^{(\mu \| Z_j - Z_i \|)}$$

where r is a random variable drawn from a uniform distribution on $[-0.3, 0.3]$. The variable Z_i is the location of the ith unit in the array of units and corresponds to the index of that cluster within the entire layer, thus the third cluster has $Z_i = 3$ for all units i in the third cluster. At present inhibitory connections are not shaped by learning.

Learning Equations. Learning of excitatory weights is Hebbian, and follows Malsburg (1973) in requiring that the sum of all excitatory weights to a unit remain constant over time. Weight normalization implements competition between incoming signals that heavily favors connections between simultaneously active units.

$$\tilde{w}_{ij}^+ = w_{ij}^+ + \epsilon x_i {x_j}^2$$

$$w_{ij}^+ = E \frac{\tilde{w}_{ij}^+}{\sum_i^N \tilde{w}_{ij}}$$

The network learns on every time cycle. Over time the synaptic weights encode the spatio-temporal correlations that occur when delayed lateral excitation is strongly correlated with bottom-up activation from input to the motion detection layer. Because shorter connections are initially stronger, units are more likely to encode local transition information.

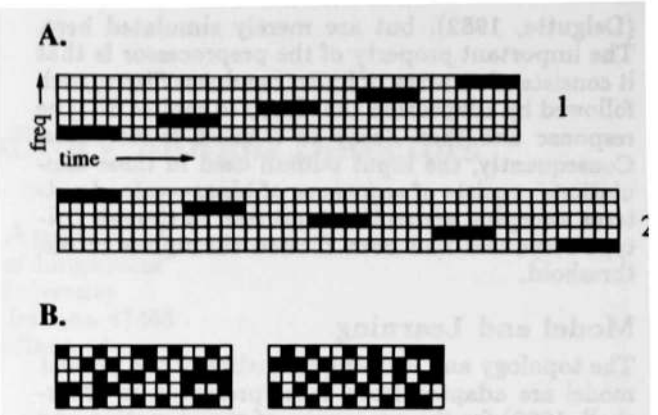

Figure 2: Artificially produced stimuli used to examine self-organization of motion detection. **A.** Frequency modulated "up" and "down" sweeps at two different rates (1/8 and 1/10). **B.** Bursts consisting of random input for limited durations.

Simulations

FM Sweep Stimuli. The self-organising properties of the model were studied in conditions corresponding to ideal realizations of input from cochlear preprocessing. A network consisting of 10 input units and 10 clusters of 3 motion detection units was exposed to FM sweeps beginning at all 10 of the units (i.e., all ten different frequencies) for 10,000 time steps. Monotonically increasing and decreasing sweeps occurred at 3 different rates (1 frequency step per 8, 9, and 10 time steps). The input for each frequency simulated ON-type cell response by remaining on for 5 time steps and then falling to 0. Spectral representations of some of these stimuli are shown in Figure 2. The inclusion of stimuli that begin at all frequencies enhances learning of units along the edges of the motion detection layer by reinforcing delayed connections from tonotopically near neighbors. If stimuli begin only at the edge of a detection layer, distant connections are most relevant to detection at the other edge and motion is not disambiguated as well for those units. All stimuli were separated by periods of zero input to permit previous activations to decay. In the absence of input most activations decayed after about 3 iterations.

The values of parameters used in all simulations are listed below.

Description	Parameter	Value
discretisation	Δt	0.2
unit time const.	τ	3.3
unit decay const.	γ	1.1
sum of weights	E	1.3
lateral delay	k	10
learning rate	ϵ	0.07

As would be expected, those networks exposed to stimuli at a single rate (not shown) develop the most discriminative code. The presence of stimuli at other rates blurs the temporal correlations arriving at successive units in the motion detection layer, but motion detection is quite robust across the three different rates. The output of input and some of the output units is shown in Figure 3 for upward and downward sweeps at the fastest and slowest rates. For the sake of clarity only a subset of the motion detection outputs are shown, although the units not shown here responded similarly. In the figure the 10 input units are shown at the bottom of the graph, while the last 4 clusters of units are grouped and drawn above maintaining their tonotopic relationship. The first 2 sequences exhibit upward FM sweeps, whereas the latter 2 exhibit downward sweeps. Dashed lines have been placed at activation values of 0.7 to permit comparison of unit activities. Consider downward sweep first. Note that units 28-30, the first units to fire for downward sweeps, are stimulated only by bottom-up activation and therefore remain moderately active but do not show winner-take-all behavior. Later, unit 27 of the next cluster of units (25-27) and then unit 20 of (19-21) receive both bottom-up and time-delayed lateral activation and thus go supra-threshold, consistently encoding direction of motion for downward sweeps at all rates. In like manner, unit 25 encodes direction of motion for upward sweeps. Disjoint subsets of units in each cluster learn to encode the two possible directions, though the cluster (22-24) does not respond well to downward sweeps. Finally, note that the response to upward sweeps is both greater in value and longer in duration. This occurs because later firing units receive input from a larger set of coherent motion detection cells already responding to direction of motion.

Bursts. It is extremely important that motion detection learning be robust despite the introduction of noise and constant frequency components, since both types of stimuli are well represented in natural environments. We did not examine constant frequency stimuli, since these involve self-excitation of one cluster and therefore produce no correlation between bottom-up excitation caused by spectral motion and lateral excitation patterns. However, the effect of bursts like those shown in Figure 2, which produce spurious correlations, were simulated. When noise bursts of duration 5, 7, and 10 (random input) were added to the FM task outlined above, the motion detection layer still reliably encoded direction of motion at all rates.

Discussion

There is an important relationship between stimulus duration and the constant k that determines the duration of transmission delay. If stimulus duration approaches the value of k, distant units may be simultaneously active, leading to ambiguity in the direction of motion. This can cause incorrect learning or, worse, the development of weights that cause some units of the network to permanently saturate. Thus, stimulus duration must be sufficient to permit competition between units in a cluster, but must not be so great as to cause too many simultaneously active units in the motion detection layer.

It is instructive to compare this formulation of feature processing with that which is implied by bottom-up time-delay systems. If one were to assume that motion sensitivity of the sort advocated in this report were founded on bottom-up time delays, the motion detection layer would necessarily have to map dissimilar inputs to the same output (See Figure 4.) This problem arises because as a time-delayed pattern sweeps across L2, its manifestations at different points in time are entirely unrelated. In Figure 4 the same input pattern at two successive points in time is labelled $P(t-1)$ and $P(t)$. As Rumelhart and McClelland (1986) note, solutions to this problem can be found by incorporating a hidden layer of units. Unfortunately, in this case a hidden layer of units leads to a very abstract, non-tonotopic code for motion that is not easily learned without some form of supervision. These problems are overcome in a very simple manner if bottom-up time delays are replaced by lateral delays that permit the learning of spatiotemporal correlations.

This report shows how the shunting equations used by Marshall (1990) can be reformulated and combined with a different learning rule to endow a network layer with the ability to encode direction of motion. The motion detectors arise as chains of active units in response to statistical regularities that would occur over a 1-dimensional tonotopic array of units with receptive fields limited to a small band of frequencies. Members of the chains of motion detectors that arise for more rapid spectral patterns continue to encode direction of motion, although the code becomes spatially and temporally sparse. From the standpoint of local computational constraints, detection of auditory spectral motion provides a means for discriminating two patterns that may well excite the same group of neurons on the basis of direction of motion. Output from the motion detection network can then be interpreted by networks that learn spatial patterns, leading to more general sequential pattern recognition.

References

Aitkin, L. 1990. *Information Processing in Mammalian Auditory and Tactile Systems.* New York: Wiley.

Carpenter, G. and Grossberg, S. 1987. A massively parallel architecture for a self-organizing neu-

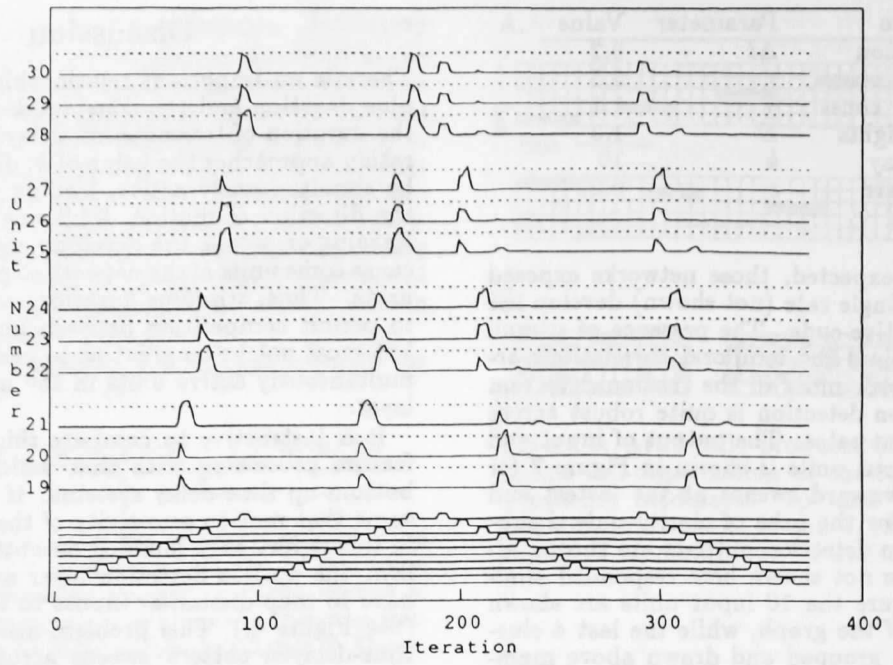

Figure 3: Unit activations over time during presentation of 4 FM sweep stimuli. The unit number of each unit in the motion detection layer is listed at the left of the graph. The first two stimuli are progressively faster upward sweeps at the rates 1/10 and 1/8; the last three stimuli show response to downward sweeps in the same order. Dashed lines indicate an activation of 0.7. Responses below this value are considered sub-threshold. Note that for each cluster, units that win the competition for one upward sweep also win for all other upward sweeps. (Input stimuli have been scaled for purposes of illustration.)

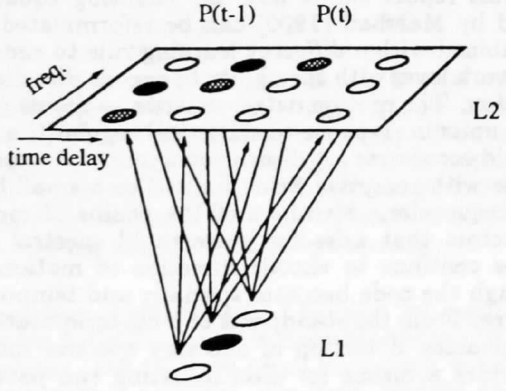

Figure 4: Pattern of activations as a pattern sweeps across L1 and leaves a time-delay trace across layer L2. Activations shown in L2 are the superimposition of two discrete time steps; the activations arising from motion of a single pattern are shown in two shades to indicate that they are not simultaneous.

ral pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 37:54–115.

Cohen, M. and Grossberg, S. 1983. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:813–825.

Delgutte, B. 1982. Some correlates of phonetic distinctions at the level of the auditory nerve. In Carlson, R. and Granstrom, B., editors, *The Representation of Speech in the Peripheral Auditory System*, pages 131–149. Elsevier Biomedical Press.

Delgutte, B. 1986. Analysis of french stop consonants using a model of the peripheral auditory system. In Perkell, J. and Klatt, D., editors, *Invariance and Variability in Speech Processes*. Hillsdale, New Jersey: Erlbaum Associates.

Elman, J. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

Földiák, P. 1991. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

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

Kohonen, T. 1984. *Self-Organization and Associative Memory*. New York: Springer-Verlag.

Malsburg, C. 1973. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100.

Marshall, J. 1990. Self-organising neural networks for perception of visual motion. *Neural Networks*, 3:45–74.

Metzger, Y. and Lehmann, D. 1990. Learning temporal sequences by local synaptic changes. *Network*, 1:169–188.

Pickles, J. O. 1988. *An Introduction to The Physiology of Hearing*. New York: Academic Press. Second Edition.

Reiss, M. and Taylor, J. 1991. Storing temporal sequences. *Neural Networks*, 4:773–787.

Rumelhart, D. and McClelland, J. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of cognition*, volume 1. Cambridge, Massachusetts: MIT Press.

Suga, N. 1990. Cortical computational maps for auditory imaging. *Neural Networks*, 3:3–21.

Wang, D. L. and Arbib, M. 1990. Complex temporal sequence learning based on short-term memory. *Proceedings of the IEEE*, 78:1536–1542.

Whitfield, I. C. and Evans, E. F. 1965. Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology*, 28:655–672.