Toward Audition in an Open Environment[1]

Robert F. Port **Sven E. Anderson** **J. Devin McAuley**
Program in Cognitive Science
Indiana University
Bloomington, Indiana 47405

[1]This work was supported by the Office of Naval Research Grant No. N00014-91-J1261, the Indiana University Graduate School, the Armed Forces Communications and Electronics Association, and by the National Institute of Mental Health

1 Introduction

Our research program is directed toward understanding human auditory cognition by means of simulated auditory cognition. That is, we seek to construct functional models that actually hear in a biologically plausible way. In our view this approach will be more fruitful than, for example, focussing on 'speech recognition' or 'pattern recognition' or simulating discrimination tasks. If we know one thing about the auditory systems of adult higher animals, it is that they are highly customized to a specific acoustic environment – the range of environments in which the animal lives. On the other hand, any animal can also adapt to a novel environment. It can learn to ignore prominent sounds that have little practical consequence and to pay close attention to acoustic subtleties that happen to have important consequences. Thus, we propose that a major goal of audition research be to model auditory systems capable of functioning in an open environment.

By an *open environment* we mean, not one that has no constraints whatever regarding what can happen next (like say sunspots or cosmis rays), but rather an environment in which there are many events and subevents that recur with various probabilities, and in which only some of these events are time-linked to each other. That is, an open acoustic environment has acoustic events that recur with a range of statistical properties. And these distributions themselves may change slowly over time. Thus we propose that an important long-term goal of research on hearing should be to develop a simulated auditory system that can adapt to an acoustic environment by recognizing the statistical regularities in stimulation. One problem with such an approach is that the very open-endedness of the task makes evaluation of the model difficult. Evaluation of the performance of any model must be done by sampling from an open set of possible acoustic environments. This can be handled by constructing an environment-generating system.

In this chapter we will discuss some general properties of audition in an open environment emphasizing the constraints imposed by timing on this process and reviewing some previous attempts to model temporal patterns. Then we will discuss two models developed in our lab that will be form useful components of an auditory system capable of dealing with an open environment.

Auditory Environments. Perceptual systems should be studied within some environment since the primary function of an auditory system must be to extract information for its 'owner' about events in that environment. This focus differs from more traditional attempts to do mere pattern recognition. To recognize patterns implies differentiating them from competing non-target patterns (and not with high signal/noise ratios). In our acoustically cluttered world, the statistical properties of the noise are as important as the statistical properties of 'targets', and recognition under high signal/noise ratios is an ideal that will occur only infrequently.

What are the general properties of an acoustic environment? First, since sound waves interact as a three-dimensional linear system, the sound generated by independent events occurring in different places are simply superimposed on each other. Thus, separating independent sound sources is a major problem (Bregman, 1990). In our research thus far, we have developed a fairly robust sequence detecting system that can adapt itself to an simple auditory environment, that is to a field of competing sounds plus white noise.

A second major feature is that many environmental events are periodic but have periods that change from period to period. Thus an effective auditory system needs to exploit the regularity despite the change of rate by 'locking in' to periodically repeating events. We have done work on an adaptive oscillator that responds appropriately to periodic inputs. An effective auditory system will be able to recognize periodic sounds of any shape - including periodic squeeks, chirps, tweets, etc. Competing periodicities of similar events will also need to be separated.

A third critical feature of an acoustic environment is that (a) it may change slowly over time and (b) the environment will be different in different places that an organism might locomote to. Thus an auditory system for an open environment must be able to continuously adapt to changing statistical properties of the context.

Our purpose in this paper is to consider some general features of audition in higher animals, human or otherwise, and to describe several simulated components we have developed in our lab that we intend to serve as parts of an auditory system for such open environments. But what should such a system be able to do? Obviously that depends on many details about the animal and its environment. But with some careful thought we can propose some general specifications.

First, we sketch some major specifications of a functional system – a model that will adapt to recognize salient sound patterns in some range of acoustically complex environments. We propose the following features:

1. Real Time Performance. A suitable model should run in real time. It does not appear to be too difficult to run a system fast enough for real-time interaction on essentially any modern work-station if suitable DSP support is available.

2. Continuous Adaptation. A model should be adapt to slowly changing properties of the stimulus patterning. Of course, this features has a 'downside': the model will be able to cope well in the short run only with environments with which it is familiar. This property certainly agrees with human performance.

3. Spectral Description. As an initial step, any system needs to reduced high-frequency components to a spatial code – that is, do mechanical frequency analysis.

4. Time-Varying Patterns. Many patterns take time to manifest themselves. A functional auditory model should be to pick these up too, no matter how they are defined.

5. Binaural Localization. Though we do not address this issue at all, an effective auditory processing system should make use of binaural information to locate sound sources in space.

6. Source Segregation. An auditory system should be able to segregate superimposed acoustic events that represent different environmental events. This will, as pointed out by Bregman, come in two forms: one kind of segregation is based on familiarity with the patterns and the second is based on more raw acoustic features (reflecting, no doubt, statistical regularities in the stimulus field).

This implies self-organized learning and the ability to ignore structured noise as well as to ignore modest rate variation.

In this paper, we also present our current models that address the recognition of auditory patterns that are defined by how sound changes over time. Our focus is on complex non-speech auditory patterns that range in duration from about 1/10 sec up to several seconds, plus periodic series of such patterns.

In order to focus on the issue of how time will structure the recognition of auditory patterns, we will sidestep in this chapter a number of other significant aspects of audition in higher animals. First, we assume without further discussion the existence of a spectral analysis mechanism which analyzes the temporal microstructure of the acoustic waveform into parallel spectral channels. This is roughly equivalent to assuming an 'ear' and taking our input to be a stylized version of the auditory nerve representation of sound. Second, we leave aside the many issues that arise from the existence of two ears: binaural processing, spatial localization of sound sources, channel separation, etc.

We have sought whatever insights we could glean from neuroscience research about how animal auditory systems process temporal information and used them to guide us in designing an architecture that is suitable for the identification and discrimination tasks. For example, there are spatial layouts of neurons based on frequency at several different level of the auditory system. Nevertheless, we have forged ahead with what might be characterized as a "biologically-inspired engineering approach" to see what might work. Thus, the role of biology in this program of research has been both to help us formulate the problems that need to be solved by an auditory system functioning in a noisy environment, and also to determine the specific design features of the system. Where useful we also employ methods of experimental psychology to evaluate and guide the model development by comparing model performance to relevant behavioral data.

The structure of this paper is to first consider the major classes of temporal information that are important for organisms, then to review previous attempts to use artificial neural networks to recognize sequential patterns. Then we describe two models developed in our lab that will together form the backbone of a general auditory pattern processing system. The first is the Lexin model for recognition of complex sequences (Anderson, 1994) and the second is a framework for design of adaptive oscillators (McAuley, 1994a) that will serve as a component of a more general model for recognition of periodic patterns of many kinds.

2 Temporal Aspects of Auditory Cognition

Our focus in this section is on the kind of temporal information that can arise in such an environment. But the specification of any kind of temporal information depends on the prior specification of an inventory of Spectral Features. It seems that logically, before we discuss time, we must have shape. That is, any auditory system must differentiate some set of spectral shapes (that is, do some form of frequency analysis and note differences in the spectrum. We imagine these features as including simple spectral 'peaks' – regions of relatively greater energy or as level-independent 'spectral profiles' (Green, 1976), etc. Methods for extraction of such spectral descriptions using neurally plausible mechanisms are well known (Grossberg, 1976).

Then, given some instantaneous spectral primitives, we find it helpful to classify auditory information into three types according to the temporal pattern structure that is of ecological relevance (see Figure 1): *Acoustic Motions, Rhythms* and *Sequences.* Most sounds that display changes over time, from insect chirps and chair squeeks to human speech, fall into at least one of these classes. Each of these three types seems to imply distinct kinds of architectural mechanisms for its extraction from the signal. In this section, we will sketch some examples of each of these three types of information.

Acoustic Motion. Rate of change information is that which can be extracted by detection of a change in an instantaneous spectral feature over some arbitrary but short interval. Thus, one could detect a change in the frequency of a salient spectral peak over a brief period of time, or a change in the intensity of a frequency component over a brief period of time. This requires detection of some prominent event (eg, a spectral peak) and then detection of a local change. The mechanisms for detecting spectral change are not especially difficult to design (Smythe, 1988) and neural circuits capable of picking up this information have been discovered in the sensory systems of many animals including humans.

Detection of motion in a spectral peak over time requires several things, first, the detec-

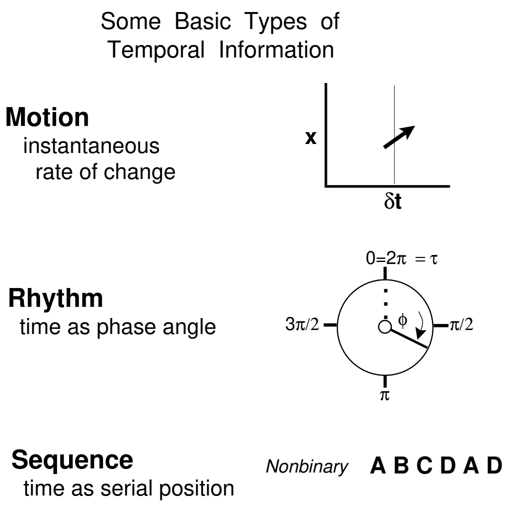

Figure 1: Three basic types of temporal information important for biological systems.

tion of a spectral description, and then, either inhibition or excitation of the 'same' feature detectors in neighboring frequency regions after a brief time delay (?; Anderson, 1994). For example, using cells with variable time delays, feature detectors for rising and falling peaks can be constructed. In the case of speech signals, the slopes of the frequency change in the spectral peaks are known to carry important information about speaking rate and about the classification of a speech sound as a consonant or a vowel.[2] Just as the static spectral features provide the input to a dynamic auditory feature, such as a *tweet* or a *chewp*, the *tweet* and *chewp* features, in turn provide the vocabulary for specification of higher-level and longer-duration spectral patterns – at least if we design it correctly.

Rhythm. Recognition of a rhythm is a more complex problem. A minimal rhythm can be defined as a series of repeated similar events. When such a series of regular prominent events occurs (let's call them the 'primary series' that define zero phase angle), other, nonprimary events can be localized in time by measurement of their phase angle with respect to the

[2]It has also been proposed that the change in the rate-of-change of intensity in the sound that is reflected from an approaching object provides an "auditory tau" which directly represents time-to-contact – a measure that is independent of the absolute rate of approach to the object (?; ?).

primary series. Of course, musical notation (using a time signature, quarter notes and half notes, etc) is a notational implementation of the notion of phase angle. Rhythmic events arise ecologically due to footfalls, swaying branches, swinging animal limbs, bouncing objects, etc. In general, they can result whenever large physical objects oscillate mechanically. When a complex event occurs, such as the periodic production of footfalls in quadraped locomotion, the relative durations of certain internal events can be measured by how much phase angle they occupy with respect to an underlying beat duration.[3]

While listening to a horse, one can tell a trot from a gallop (given the same period) by the phase-angle pattern of the footfalls. Similarly, a waltz differs from a foxtrot or a tango by the number and phase angles of the beats that lie within the primary metrical cycle. Similarly, engines tend to produce rhythmically structured patterns that may vary (irrelevantly) in rate.

A mechanism that can recognize a periodic event must be interpretable as kind of resonator - a device that 'rings' or responds strongly at a particular frequency. One simple device that can recognize simple periodic events would be a set of pretuned oscillators (like, for example, the unstopped strings of a piano), but to recognize the internal phase angles of events within the metrical period, independent oscillators will not do the job by themselves. A simple acoustic resonator can only 'recognize' a single kind of input events series; a change in frequency in the pattern means that a new resonator exhibits sensitivity. Instead, a system built of tunable oscillators is more useful – especially if it lies upstream from detectors for other more primitive events. Such a device can tune in to a pattern at one rate and then remain tuned to it as the rate changes (at least if it does so slowly).

A mechanism for such an adaptive oscillator will be proposed below. To stick with a simple example (but one that cannot be handled by a simple bank of filters), let us imagine detecting the presence of, say, a repeated series of rising glissandos (something like '*tweet, tweet, tweet...*') which need to be distinguished from a competing sequence of falling glissandos (like '*chewp, chewp, chewp...*'). To track the regular *tweets* and not be confused by

[3]Presumably this is how 'long' and 'short' vowels are distinguished in human languages (like, eg, Finnish and Arabic) despite changes in speaking rate. It has also been argued that, essentially, phase angle time measurement (that is, speech rhythm) plays a role in distinguishing syllable-final voiced from voiceless stops and fricatives, in English (Port and Dalby, 1982; Port and Cummins, 1992)

the competing *chewps*, one would need to first detect a single *tweet* and then feed the output of the 'tweet-detector' to an adaptive oscillator mechanism that predicts when subsequent *tweets* will occur. The output of this oscillator, integrated over time, is the signal indicating that a series of *tweets* is under way.

Sequences. When most people think of 'temporal pattern recognition', they think automatically of sequence recognition. Linguists, for example, have traditionally viewed the problem of speech and language perception as essentially a sequence problem. Words and phonemes are assumed to come in sequence just like letters on a page. Similarly, the sequencing of just two types of basic events, 0s and 1s is the essence of digital communication. Yet ordered events actually comprise the most abstract kind of time, since it is the form of temporal measurement to which the most powerful distorting transformations can be applied without significant effects (Port, 1986). For example, if one is only interested in the order of ABC, then the duration of the event called A and the temporal spacing between A and B is ignored – by definition. Although sequence may seem the simplest kind of time, it actually requires very powerful processing mechanisms if sequence recognition is to be implemented with appropriate generality.

In general, differentiating one sequence from another requires saving up either the raw input signal, some description of it (perhaps including a rate-invariant description) of sufficient duration to recognize the whole pattern. Thus the target patterns themselves determine the amount of memory required. The next section reviews a number of connectionist architectures that address the question of recognition and memory of sequences.

But what about the rate of presentation of the pattern? If the input sequence arrives with the pattern elements already distinguished, then the length of the pattern determines the memory requirement. But many patterns that are 'the same' may appear at a range of rates. This is true of human speech, familiar footsteps, and many other periodic sounds including music. Recognition of such patterns presents serious difficulties since the amount and form of memory required in absolute time seems to depend on the rate of presentation of the pattern. If memory is to be measured in terms of patterns elements rather than in absolute time, then the system must be able to recognize each pattern element as it arrives – no simple matter. Thus, sequence recognition in real time poses many fundamental problems.

These three kinds of temporal pattern information suggest different kinds of recognition mechanisms. It seems to us that most temporal information used by auditory systems falls into one of these three types. Acoustic motion detectors are the simplest to design – requiring only time-delays linking neighboring frequency detectors. Processing units in the first layers of the Lexin model will be shown to develop sensitivity to acoustic motion.

Sequence recognition is a challenging problem because it permits a great degree of complexity and memory in the inventory of patterns that are possible. However, this is inevitably bound up with the actual inventory of patterns that *need* to be recognized. Work on the problem of sequence recognition has led to the Lexin model which is capable of developing units sensitive to a wide range of complex sequences that occur frequently in the surrounding environment of patterns.

For rhythm detection, we assume a Lexin-like front-end that is first able to identify salient acoustic features. An entrainment mechanism, embodied as an *adaptive oscillator* will be shown to lock-on to the semi-periodic occurance of an onset representation of pre-identified features, thereby enabling the system to predict and track the future occurance of rhythmic features in noise. Recent work in our lab has been directed toward developing this entrainment based model of rhythm detection. Investigations with the adaptive oscillator model have been so far limited to simple rhythms.

In the next section, we will review and comment on a number of models developed within the tradition of artificial neural networks for the recognition of sequential patterns. This problem is typically formulated far more generally than our biological orientation allows. The problem presents many difficulties and raises very basic conceptual issues for cognitive science such as 'what do we mean by memory?' and 'what does our system need to know this information for anyway?'

3 Network Approaches to Sequence

Much work has been done on the sequential aspect of temporal information using connectionist networks. Although some of this work addressed sound (eg, in speech recognition), there has been little understanding in the literature of the ways in which temporal patterns differ from mere sequential order. We review some of these attempts at sequence recognition using neural networks and will comment on ways in which they are inappropriate for the kinds of temporal information used by organisms dealing with an acoustic environment.

If networks are to address sequential patterns, there are three temporal tasks a network might be trained to perform: *sequence recognition*, *sequence reproduction*, and *temporal association* (Hertz et al., 1991). For the reproduction of a sequence, the network is required to finish a partially presented sequential pattern. Sequence recognition requires the network to generate a response to a sequential pattern. The response itself may not be sequential. By contrast, temporal association requires a network to produce a particular sequence in response to a particular sequential input, perhaps entraining movement of parts of an organism to some sound pattern.

To recognize sequence a nervous system must first cast the stream of environmental stimulation into a form suitable for neural processing. Because neural networks are spatially high-dimensional, they are quite well suited to spatially distributed processing. In fact, many artificial networks employ static data structures that cast time-sequence data into large spatial input arrays. Turning time into space is a common theme among those who have attempted to deal with time. Spatial input vectors at times t_1 and t_2 can then be mapped to a single set of spatial arrayed output vectors, as long as an input at t_1 does not need to affect the processing of some later input at t_2. In our view, this method of dealing with time is not very useful for pattern recognition (Port, 1990) although it surely is useful for spatial localization with binaural inputs (?; ?).

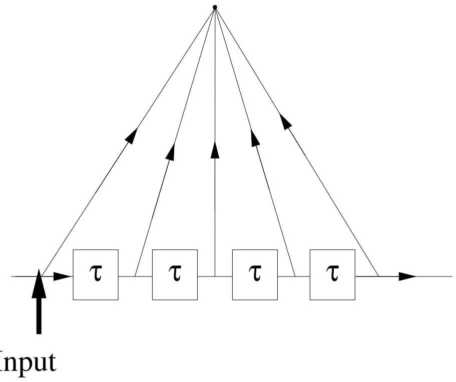

Figure 2: Delay line memory. A delay line tapped at various positions transforms temporal data into a spatial array.

3.1 Mozer's Typology of Memory Mechanisms

Mozer (1993) recently proposed basing a taxonomy of network architectures for temporal pattern processing on three dimensions by which memory in networks can be described. We should consider whether his typology will suit our needs. It our opinion it does not. His three dimensions are the *form*, *content*, and *adaptability* of memory. The form of memory is the mechanism by which information is retained from one moment to the next (assuming the discretely sampled time units, of course). The content of memory is some transformation of the network's input and/or state. Adaptability simply refers to whether parameters important to memory, like weights and biases, are constant or learned.

For example, Mozer's simplest form of memory employs a buffer to hold the most recent inputs. A so-called 'delay line' serves this purpose. A delay line is a series of nodes spanning some duration along which input signals move at each time step to create a spatial array representing the recent series of input, as shown in Figure 2. Each new input causes all the others to move down one step, and the oldest stored input is lost from the array.

Mozer points out that a huge variety of forms of memory can be formally realized via the convolution of the vectors to be remembered with an arbitrary kernel function $c_i(t)$

$$\bar{\mathbf{x}}_{\mathbf{i}}(\mathbf{t}) = \sum_{\tau=\mathbf{1}}^{\mathbf{t}} \mathbf{c}_{\mathbf{i}}(\mathbf{t} - \tau)\mathbf{x}(\tau)$$

where the $\mathbf{x}(\tau)$ are the input vectors to be remembered (Mozer, 1993). This formulation isolates the mechanism that links time steps, and accommodates delay-line memory as well as a memory composed of exponentially decaying unit activations. The recursive update equations for unit activations combine time delays and self-recurrent terms. While Mozer points out that "any kernel function results in a distinct memory form," this is not, in our view, the most useful way to classify memory systems that are used to recognize familiar patterns.

For the purposes of describing auditory temporal processing models, the distinction between form and content seems less important than the distinction between different possible kinds of memory contents. First, the 'form of memory' is a rather vague notion, since most recurrent networks employ a combination of delays (often simply implicit delays between layers) and decaying activation or changing network states. Second, the concept of 'form' seems to group dissimilar memory types. Consider a memory that is maintained by virtue of the state the network is presently in. Zipser (1991) has proposed a memory in which stored values on one unit in a completely recurrent network decay to a fixed point of the network, much like single activations decay in exponential memory. This "state" memory is specified by the contents of memory—the network's state , and maintained by many units with complex behaviors. Functionally, this type of memory is very similar to exponential trace memory. However, it cannot be correctly described by a convolution of the input with a kernel function of the form $(1 - \mu_i)\mu_i^t$. The formal implementation of the memory seems to us to be less important than the information it ultimately encodes.

Finally, there is the question of exactly what "memory" means. For the recognition of completely arbitrary sequential information, memory must retain as much detail as possible about an input sequence. If, on the other hand, a system is designed to recognize familiar sequential patterns, there is no need for memory to be completely general purpose. To the extent that auditory pattern recognition is intended to mean the recognition of *familiar* patterns (as is the case with typical biological systems in a slowly changing environment), short-term memory need only retain enough information to discriminate among the different familiar patterns. It is not irrelevant that Mozer's paper appeared in a book dealing with the problem of predicting a completely unknown sequence. Our primary interest, as noted above, is in the temporal pattern recognition problems faced by organisms living in an open set of ecosystems. Whereas the content of memory may be related to the set of patterns that must be learned, the form of memory is based on general principles that describe how activations persist over time.

For these reasons, we believe auditory temporal pattern recognition networks are best classified according to the *contents of memory*, according to what information is stored about he past. From this point of view, the simplest type of memory, then, is a simple *copy of some number of past inputs* as provided by a delay line. The raw signal is simply saved veridically. More complex types of memory depend on inputs that are *transformed by monotonic functions*. In this case, a value specifies how recently an input was entered. Other transforms of input are generally more complex to interpret. Finally, the most complex types are *state memories*. These depend on the previous contents of memory as well as on inputs so that the activation state of the system, representing a context sensitive analysis of what has occurred, is the memory. Thus, there is little or no 'representation' of sound preceding actual 'recognition' of sound. In our view, architectures of this class will eventually prove most useulf for an open environment – and will also prove to model human perception more closely.

3.2 Review of Sequence Recognition Models

The next three subsections describe a number of network models of sequential pattern recognition. Since the focus of this discussion is on memory mechanisms, the models are grouped according to the complexity of short-term memory mechanisms employed by each model, moving from memory of raw inputs to memory distributed over entire networks.

3.2.1 Memory for Inputs

The most common models of sequence recognition have employed a temporal window or set of delay lines large enough to hold an entire input pattern (Kohonen, 1984; McClelland and Elman, 1986; Sejnowski and Rosenberg, 1987; Shamma, 1987; Waibel et al., 1987; Elman and Zipser, 1988). Once sequential patterns have been transformed to spatial patterns, any one of the numerous spatial pattern associators can be employed for sequence recognition.

Perhaps the most sophisticated example of a network that retains a record of past input patterns is the *time-delay neural network*, or TDNN, developed for speech recognition (Waibel et al., 1987; Lang et al., 1990). The network is strictly feed-forward and has two layers of units connected by feed-forward connections. Units in the first layer receive time-delayed inputs. Second-layer units receive inputs from the first layer that are transmitted via connections having even longer time delays. These networks are trained using a supervised algorithm by adjusting weights to minimize an objective function.

A major drawback of this form of memory is that the input must be clocked with respect to the iterations of the network by some agent that is outside the network itself. This makes it implausible as a neural model. Furthermore, it postulates that the raw auditory signal is stored up only long enough to permit recognition of a targeted set of auditory stimuli. What about similar *non-target* stimuli that vary in their rate of presentation, as will arise in an open environment? Overall, this method is brittle and depends on a fast biological clock for syncronizing events at different locations in the system. Thus is it is implausible.

3.2.2 Memory for Transformed Input

The next simplest memories consist of transformed inputs. These usually result from activation decay. Tank and Hopfield (1987) use multiple delay functions for each input to facilitate recognition of temporally distorted sequences. The delay functions broaden the duration of inputs using averages of past input values. Essentially, longer delays are associated with temporally broader input distributions. Units that recognize patterns receive their own feedback in addition to the delayed input, making this a hybrid approach that uses both direct and transformed inputs. Networks that use this delay mechanism are less sensitive to minor sequence distortions, and thus solve much of the clocking problem discussed above (Hertz et al., 1991). On the other hand, exponential decay of raw inputs does not provide a very reliable memory for past.

Wang and Arbib Model. Wang (1990) has demonstrated that a model that maintains temporal memory via a decaying spatial pattern of activation can be used as the basis for recognition of complex sequences [4]. The model consists of a layer of short-term memory units and a layer of recognition units. Memory and pattern association are maintained in separated sets of units. Initially, recognition units receive connections from all memory units. Short-term memory in the model is provided by recurrent connections between dual units. The units form a damped oscillator with activation peak values that decay exponentially. Spatial pattern association then takes place over a large number of these memory units. A unit in the recognition layer becomes selective for a sequence if that unit is chosen to become activated at the end of the sequence. The learning algorithm thus requires specification of a unit and a point in time at which sequence recognition is to take place. The learning process gradually moves the vector representing weights to a recognition unit until it is parallel to the trace of activation in short-term memory. Since weights are normalized, the chosen unit responds optimally to the selected memory pattern. If the unit's threshold is chosen appropriately, it will be the only unit excited by the sequential pattern. Recognition of complex sequences in which one or more components may be repeated the network requires the addition of a (biologically implausible) stack of units that shift previous activations when an identical token occurs. One strength of the model is its ability to learn arbitrarily complex sequences. In addition, once units detect a subsequence they can anticipate and prime units that recognize later elements of a sequence.

Riess and Taylor. A similar use of activation decay for memory is developed by Reiss and Taylor (1991) for sequence storage and completion. In this model, pattern association and memory are also maintained by distinct parts of the architecture. In addition to variable activation decay rates, memory units require a specified interval to saturate. Individual units

[4]Several authors characterize sequence complexity in terms of the *degree* of a sequence. The degree of a sequence is intended to be the minimum number of preceding components necessary to correctly predict its occurrence (Dehaene et al., 1987; Wang and Arbib, 1990). For instance, the symbol **8** in a set of sequences comprised of the single sequence **3-1-2-1-8-1-7-1-2-1-9-4-3-2** is of degree 4, because to correctly predict the occurrence of component **8** the component preceding the subsequence **1-2-1** must be known. A total of four prior components are necessary to predict **8** rather than **9** . The degree of a whole sequence is defined to be the maximum degree of all of its components. In the example, the sequence is of degree four. A complex sequence is defined to be any sequence with degree greater than one.

with a large range of decay and attack rates are connected to another set of units used to predict the next input. Decay unit activations are thresholded so that they fire with unit intensity when total activation exceeds some value. As a result only binary memory values are maintained. However, because internal variables maintain an exponential trace of previous activation, the duration of suprathreshold response can be used to indicate stimulus duration. This network reproduces sequences with the correct pattern of component durations.

Because of limitations on pattern association in a single layer, the "attack rates [are] set such that the neurons in the layer L only become active towards the end of the duration of a pattern." This effectively associates the contents of memory after presentation of one component with the trace of the next input. No features per se are recognized, although as the network predicts the next input an implicit pattern recognition occurs. It has also been shown that the decay mechanism can add contextual sensitivity to Kohonen's (1984) pattern associator (Chappell and Taylor, 1993).

Other Approaches. Several other authors have employed memory local to particular units (Anderson et al., 1988; Watrous et al., 1990; Kurogi, 1991) that retain an exponential trace of previous stimulation. The principal problem with this form of memory is that the memory-trace units combine all temporal cues, often blurring important ones with those of lesser importance. Because the models lack the structure necessary to differentiate memory occurring at different times, the activation trace of a unit does not accurately encode past events; instead, it combines them.

3.2.3 Network State Memory

In networks having recurrent connections between units, the contents of memory are derived from inputs combined with transformations of previous states of the network. The state of the network, defined at each point in time by its vector of activations, can be sufficiently unique to be associated with a particular sequence of inputs. This can be equivalent to a simple exponential trace memory (Zipser, 1991), a dynamic remapping of the input sequence (Kleinfeld and Sompolinsky, 1988), or a state arrived at only after the network makes a particular set of transitions (Elman, 1989).

Dehaene's Model. A simple example of state memory is provided by a network composed of linked synaptic triads (Dehaene et al., 1987). A synaptic triad consists of three units: unit **A** causes postsynaptic unit **B** to fire, just in case unit **C** has recently fired. If unit **C** is stimulated prior to unit **A**, then unit **B** will fire. Thus, when the network is in any state with *sequence-selective* unit **B** active, that state represents the fact that the sequence **C-A** has occurred. The network learns by beginning with a large number of synaptic triads. Those triads that are activated by input sequences have their connections strengthened. Other connections decay to zero. The resultant network can reproduce sequences or passively recognize them. No analysis of patterns occurs, and triads are not shared by different sequences. That is, two similar sequences, by virtue of having shared features, are represented distinctly by the model architecture. Thus, for a sequence to be learned it must already exist within the network; the learning algorithm merely serves to select it.

Autoassociative Network Models. A more complex state memory is based on extensions of the discrete autoassociative network (McCulloch and Pitts, 1943; Amari, 1972; Hopfield, 1982) by Kleinfeld et al. (Kleinfeld, 1986; Sompolinsky and Kanter, 1986; Kleinfeld and Sompolinsky, 1988). The network consists of a large number of binary-valued units that settle to equilibrium states, or memories. The network also makes autonomous transitions to other memories and transitions can be induced by input data.

Fully Recurrent Networks. Many researchers have studied highly recurrent connectionist networks trained using supervised gradient descent (Jordan, 1986; Port and Anderson, 1989; Elman, 1990; Robinson and Fallside, 1991). Individual units in these networks function discretely and do not retain vestiges of previous activation; memory is principally maintained by network state (Elman, 1990). These states may correspond to equilibria of the system, but they are in general extremely difficult to characterize (Cummins et al., 1993). Many of these representations are highly dependent on the set of target input sequences and do not generalize, as one might expect, to similar non-target patterns.

Memory based on network state works by interleaving processing with memory. In such a system, no direct representation remains of the input stimulus. This type of network memory can support stimulus recognition at the cost of abandoning veridical memory of stimulus details. One advantage of these systems is the on-line analysis of incoming stimuli. This permits them to recognize sequential patterns as soon as compelling evidence has accumulated. The evidence which determines sequence recognition is only weighted with respect to the competing target patterns in the systems environment. Exact memory for stimulus details is thus required only the extent it accomplishes this purpose.

Of those network models that maintain dynamic memory, very few are self-organizing. A notable exception is the work of Metzger and Lehmann (1990). This model cannot recognize complex sequences in which a component may occur in several contexts. The network does not maintain the topography of the inputs, as is found at many levels of the auditory system, and only passively recognizes sequences when it mimics transitions between learned states.

Sequence identification relies on mechanisms of short-term memory and pattern recognition spanning spatial and temporal scales beyond the receptive field of a single unit. The core assumptions in neural network modeling are that units have access only to local information that arrives from other units to which they are connected, and that the information transmitted by a unit is a fairly simple quantity. From the local perspective of a single unit, adaptation algorithms must determine the spatial and temporal receptive field of a unit. How then do local learning rules result in selectivity to afferents that are non-local? In recognizing purely spatial patterns like static patterns, a number of researchers have shown that spatial patterns can be imprinted on dense spatial arrays of neural connection strengths. The identification of real-world stimuli that are in constant flux requires that a temporal pattern somehow be transformed onto the spatial mechanism of connections combined with the flux of the network state itself.

4 Motion and Sequence: The 'Lexin' Model

In this section we describe the Lexin model for recognition of temporally complex patterns. It takes time-sampled spectra as input and employs components that detect acoustic motion to finally generate activation in units that self-organize themselves to detect complex sequences that occur frequently in the 'auditory environment' to which the system is exposed.

The model appears to be biologically plausible although many features were developed largely without specific neuroscience justification. Neural studies of the auditory system have begun to clarify how the peripheral auditory system is functionally linked to higher centers. As afferent signals travel from the ear through the midbrain to the cortex, we find that successively higher levels show sensitivity to more and more complex stimuli that represent stimulus temporal variation with less fidelity (Schreiner and Langner, 1988). As a consequence, neurons in auditory cortex are tuned to abstract features of the signal that are extremely difficult to define. Evidence from animal studies has shown that some neurons in auditory cortex respond preferentially to frequency sweeps at different rates across their receptive fields (Whitfield and Evans, 1965; Suga, 1990), while others respond to particular sequential patterns (Margoliash, 1983) or species-specific calls (Aitkin, 1990).

In songbirds many neurons in the brain are very sensitive to ordered subsets of individual components of birdsong (Margoliash, 1983). After maturation, individual neurons exhibit selectivity for artificial song-like subsequences that are frequently heard during early development. This observation is consistent with the hypothesis that neural mechanisms underlying feature discovery are plastic and develop under the influence of environmental pressures. These systems are self-organizing and thus their abilities arise without being dictated or reinforced by an external agent. The important effect of long-term auditory experience on auditory learning has also been established by auditory developmental (Gottlieb, 1975; Moore et al., 1989) and psychophysical studies (Leek and Watson, 1984).

The storage, perception, and reproduction of sequential patterns has been studied by several researchers (Dehaene et al., 1987; Kurogi, 1987; Kleinfeld and Sompolinsky, 1988; Wang and Arbib, 1990; Reiss and Taylor, 1991). Except for (Dehaene et al., 1987), few models have been offered to explain the self-organization of sequence processing. In fact, it is unlikely that the perception of sequential auditory patterns is directly tutored by and external agent. We propose a refinement and extension of the network model of auditory motion detection described in Anderson (Anderson, 1994) to general sequence identification. Sequential stimuli will consist of sequences of discrete vectors, sometimes referred to as spatial patterns, of binary digits corresponding to a firing pattern akin to that which arises in the neural place-code transduced by the basilar membrane.

4.1 Lexin Model

Recent progress in the modeling of motion detection in vision has demonstrated how a network that operates in an environment of moving, spatially coherent objects, learns to encode direction and speed of motion. The present model is qualitatively similar to Marshall (Marshall, 1990) and Anderson (Anderson, 1992; Anderson, 1994), (although single unit behavior is implemented through somewhat different mechanisms). The collection of units in which sequence-selectivity arises, the sequence identification layer (Fig. 3), 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 subnetworks that emphasize the activation of the unit with greatest activation only when that activation exceeds a certain threshold; subthreshold response is non-competitive. 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 connected with small connection strengths in a normal distribution that depends on inter-unit distance. During exposure to sequential patterns these connections organize themselves to represent the spatio-temporal correlations present between lateral and afferent excitation. A unit that receives the most simultaneous 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 in a cluster ensures that they eventually encode different patterns.

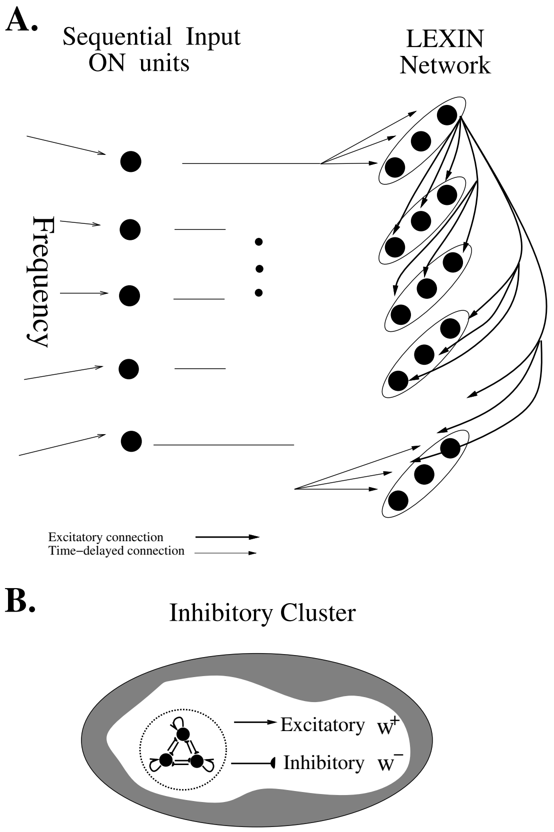

Figure 3: A. Layer of units responsible for different types of sequence detection. Actual networks usually consist of a greater number of clusters and units per cluster. B. Inhibitory clusters are enclosed by an ellipse and illustrated in the exploded view. All excitatory connections from a single unit of one cluster are shown.

4.1.1 Unit Equations and Initialization

The activation of each unit in the inhibitory clusters obeys a differential equation consisting of excitatory and inhibitory terms. Equation parameters are chosen to ensure subthreshold coactivation while providing vigorous suprathreshold competition. The selection of the most strongly activated unit in a cluster must be rapid, or intermediate activation values will impede learning.

For unit v_j in cluster C_l,

$$\tau \dot{x}_j = -\gamma x_j + f[B{x_j}^3(t) + \sum_i w^+_{ij} x_i(t-k) - \sum_{i \in C_l} w^-_{ij} {x_i}^2(t) + \Theta_j + I_j(t)], \quad (1)$$

where $I_j(t)$ is external input to v_j, Θ_j is a bias, and f is the linear threshold function,

$$f(\xi) = \begin{cases} 0 & \xi \leq 0 \\ \xi & 0 < \xi < 1 \\ 1 & \xi \geq 1. \end{cases} \tag{2}$$

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 connection strengths (weights) from unit i to j, and the w_{ij}^- are intracluster inhibitory connection strengths. Inhibitory connections were all set to one value as described in the next section. For the first two simulations reported below the value of the delay along excitatory connections was $k = 18$. In Simulation 3 delays with $k = 38$ were added. Bottom-up connections had a delay of one time step. 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.

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 and Grossberg, 1983). Unless properly chosen non-zero lateral excitatory connections will introduce positive feedback that can cause many units in the network to permanently saturate. In practice this does not occur because input units exhibit phasic response. Moreover, once a sequence has been coded, its evokes a unique response, making oscillations impossible.

Prior to adaptation excitatory connections between all units outside a cluster were set to favor connections with spatially nearby clusters. Initial weights fell off exponentially with frequency distance Inhibitory connections all had the same strength and were constant.

The behavior desired of cluster units implies a set of constraints on parameter choices sufficient to guarantee (1) decay in the absence of external stimulation, (2) competition when stimulation exceeds a threshold r, and (3) when stimulation is below a minimum threshold a. These conditions define a three-dimensional having infinite extent (see (Anderson, 1994) for details). The final choice of model parameters depends on the rate at which competition is to occur.

4.1.2 Adaptation

The model adapts to its environment by adjusting connection weights. The adaptation of excitatory connection strengths is based on competitive learning as developed in (Carpenter and Grossberg, 1987) and modified by (Marshall, 1990).

$$\dot{w}_{ij}^{+} = x_j^2(t)[(E - w_{ij}^{+})x_i(t-k) - w_{ij}^{+}\sum_{l \neq i} x_l(t-k)] \tag{3}$$

The network learns on each simulation step. During the course of learning the connection strengths 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, though this tendency may be overcome after sufficient adaptation.

It can be shown that the expectation of the sum of all connection strengths to each unit v_j approaches E and, once at that level, remains a conserved quantity. Connection strengths encode the results of unit competition as well as enhancing its effect. Thus, when a unit dominates other units in its cluster for a particular pattern, competitive learning guarantees that it will better encode this pattern on future trials. The result of the interaction of learning and unit dynamics will be clarified in simulations below.

4.2 Selectivity for Direction of Motion

The self-organization properties of the model are most apparent when it is exposed to an environment containing simple stimuli sweeping across the receptive fields of all units. A network of 10 input units connected as describe previously to 10 clusters of 3 units was presented monotonically rising and falling frequency sweeps at 3 rates (1/8, 1/9, and 1/10). Fig. 4 provides examples of two of the 6 different complete patterns. Partial sweeps, consisting of sweeps beginning at each of the input units were also presented to mitigate edge effects. Individual sweep components consisted of ON-type response for 5 time units followed by variable duration null-response.

Prior to learning no units in the network exhibited suprathreshold response to any pattern. After about 200 sweeps were randomly presented, all clusters in the network had

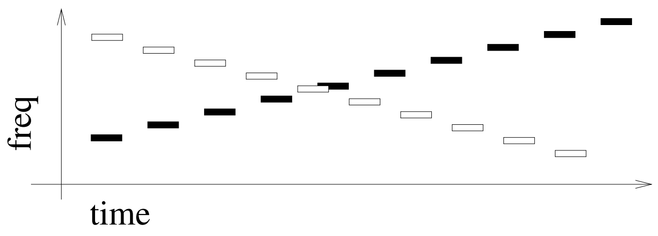

Figure 4: Examples of frequency-sweep stimuli used to induce motion sensitivity. The rising sweep has slope 1/10; the falling sweep has slope -1/9. In this case, the network was presented with only one of these at a time. Thye are superimposed for display only.

formed two units that exhibited strong selectivity for rising or falling patterns. The third unit typically remained uncommitted. The response of the first six clusters to upward and downward sweeps is shown in Fig. 5. To the right of activation traces, units have been labelled with arrows to indicate their preferred stimulus. During presentation of the first one or two sweep components all units in the corresponding cluster are weakly activated. Only after sufficient evidence accumulates do individual units dominate their clusters. For example, cluster 2 weakly identifies the rising sweep, whereas cluster 3 responds much more strongly.

After learning, the lateral connection strengths reveal how units develop complementary sensitivities to direction of motion. In Fig. 6 connection strengths are proportional to the area enclosed by boxes, the row and column indices of which correspond to the source and destination units. Tracing the pattern of activation for a rising sweep reveals how network response becomes increasingly confident as a pattern unfolds. Furthermore, it reveals how the recognition layer links activations to form a unique representation "chain." When units 1-3 receive bottom-up stimulation, they all become moderately activated and, as noted above, none wins (Fig. 5). From the weights diagram we see that the activation of units 1-3 causes unit 5 to dominate units 4-6, because only unit 5 receives appreciable excitation from units 1-3 as well as bottom-up input from the sweep itself. Units 4-6 then cause suprathreshold response in unit 8, because it is more strongly connected to units 4-6 than either unit 7 or 9. The structure of the "chaining" solution is apparent in the pattern of connection strengths in the column corresponding to unit 12. This unit shows best connectivity to unit 8, thus forming a chain of units that respond optimally to rising stimuli. A similar argument holds

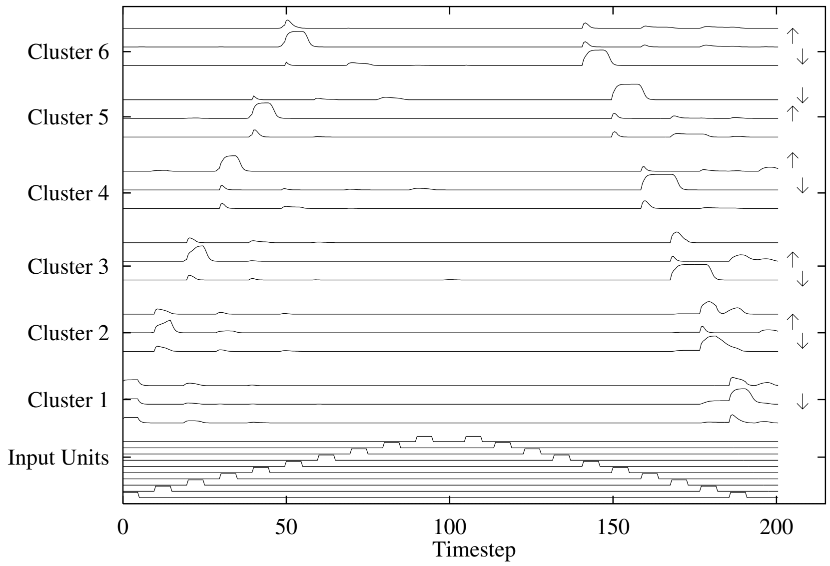

Figure 5: Time-series plot of unit activations for the first 6 (of 10) clusters of units during presentation rising and falling stimuli having slope 1/10 and -1/9, respectively. Arrows are drawn to the right of traces for units that respond selectively to rising or falling spectral motion. (Input stimuli have been scaled for purposes of illustration.)

for the response of units 14 and 17, all of which respond best to rising sweeps.

4.3 Complex Sequence Detectors

The same learning mechanisms that lead to direction sensitivity can induce a representation for more general sequential patterns. While naturally occurring auditory patterns are not arbitrary sequences, for purposes of this discussion we assume that both types of sequence can be described completely by numeric symbols representing the index of the active input unit. Additionally, it is assumed that only one input unit is active at any time.

The chaining solution to motion detection is a viable means for complex sequence recognition without using extensive short-term memory. A network of 5 input units and 5 clusters of 3 units each was presented 3 different complex sequences of degree two: **2-1-5** , **3-1-5** , and **4-1-5** . After approximately 600 random sequence presentations the response of the network to the three sequences settled to that shown in Fig. 7. Note that each of the three units in clusters 1 and 5 uniquely identify one of the three sequences.

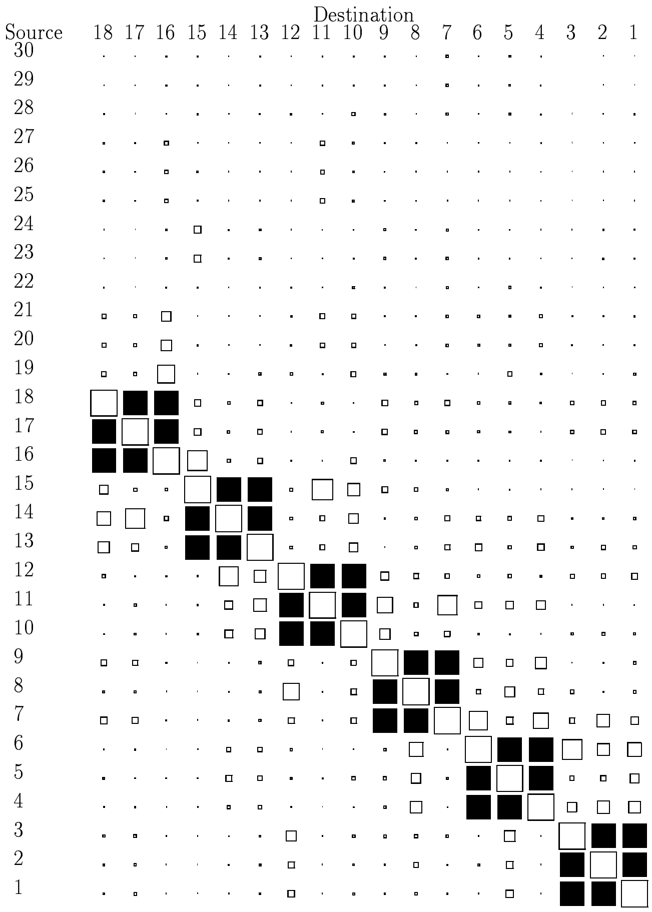

Figure 6: Lateral connection strengths. Only connections to the first 6 clusters are shown. Excitatory connections are drawn as boxes; inhibitory connections are filled. The area of each box is proportional to its magnitude and the row and column index specify the source and destination connection strength w_{ij}^{+}. Thus, each column contains a graphic display of all lateral connections to the unit labelled by that column index.

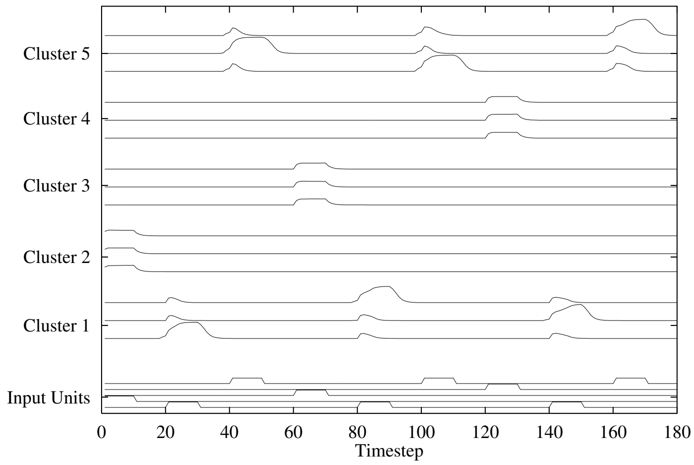

Figure 7: Response of sequence identification clusters to three sequences: **2-1-5** , **3-1-5** , and **4-1-5** (in that order). Vertical lines separate each sequence.

By induction, networks of this type can correctly identify complex sequences of arbitrary degree. Given sufficient cluster size, the chaining solution to sequence recognition requires no more memory than its longest time-delay. Complex patterns are disambiguated at the component where they first differ—thereafter the representations are nearly orthogonal.

4.4 Sequence-selective Units

An apparent shortcoming of the chaining solution is its dependence on the fixed duration delays along the lateral connections of the sequence recognition layer. When a single component is omitted, the response to a pattern may completely erode. Worse still, although strong competition is required for the early stages of adaptation, it can also cause the recognition of a complete sequence from as few as two of its components. For example, the representation of **1-2-3-4** may be indistinguishable from that caused by the input of **3-4** . There are several ways to increase contextual memory, the simplest of which involves the use of different-valued time delays. Since the competitive adaptation rule enforces the sharing of total connection strength among all connections to a unit, it is directly applicable to units linked by connections having several different delays, and implements the learning of spatio-temporal correlations at multiple time scales across multiple units.

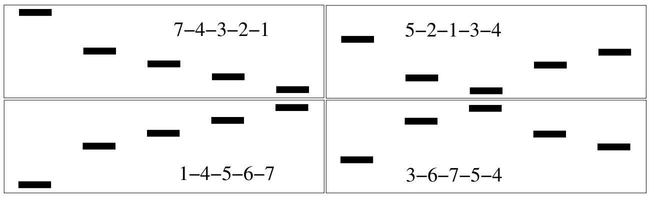

Figure 8: Four patterns used to examine unit selectivity. Like auditory patterns, they are approximately spectrally continuous. Each pattern is labelled with its numeric representation.

The enhanced contextual sensitivity of a network that uses different time delays was examined in Simulation 3. A network of 7 input units and 7 clusters of 4 units, all of which had complete sets of excitatory connections at delays of $k = 18$ and $k = 38$, was exposed to the patterns drawn in Fig. 8. The patterns in Fig. 8 are considered more realistic than those considered in the previous section because they exhibit the spectral continuity associated with sounds produced by macroscopic resonators whose continuously varying shape yields homologous spectral variation (e.g., the articulators of humans). After about 100 sequence representations most of the clusters had developed a unique representation for salient subsequences like that described for Simulations 1 and 2. Unit 19 was examined using incomplete sequences similar to sequence **5-2-1-3-4** in order to assess its selectivity. Each frame of Fig. 9 is centered over the presentation of a component (corresponding to the fourth input unit) that activates unit 19. The first frame shows the optimal final response of unit 19 following presentation of the entire sequence **5-2-1-3-4** . Frame 2 illustrates response to simple bottom-up stimulation: activation remains near 0.3—subthreshold. The partial sequence shown in the third frame, **3-4** , elicits slightly more activation which is then enhanced through competition. As more and more context is added (Frames 4 and 5) unit response nearly replicates that in Frame 1, though it is of shorter duration. In the final two frames single components, denoted by **x** were replaced by null input. In both cases response is shortened but nearly optimal. On the other hand, if the third and fourth components are *both* omitted (not shown), unit 19 response reaches a level of only 0.54, intermediate between its response to **4** and **3-4** .

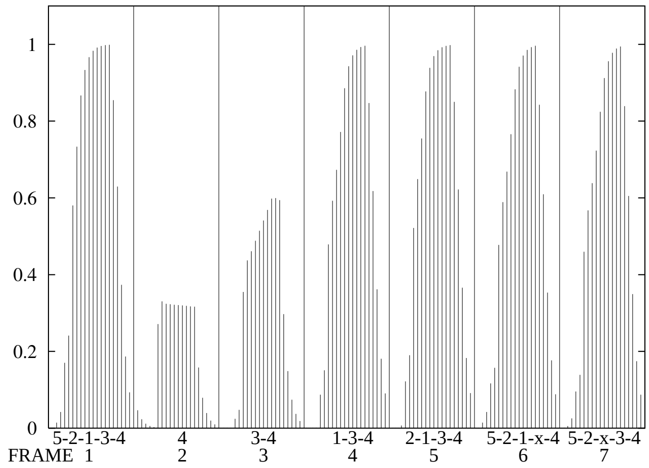

Figure 9: Response of unit 19 to various sequences ending on the bottom-up input to unit 19. Each impulse corresponds to one simulation time step; the unit activation level can be read off the ordinate axis. Below each frame the symbolic names of sequences (i.e., active units) are provided. Thus, the **4** in the frame 2 indicates that unit 19 responds at a level of about 0.3 in the absence of prior input. The letter **x** is used to denote null input for components corresponding to those time steps.

4.5 Discussion of Lexin Model

The most important feature of the proposed model is its ability to self-organize, to adapt to the requirements of different environments. Spatial patterns and temporal patterns both require the selection of an internal representation. Additionally, self-organization in the face of sequential input requires the determination of *when* a series of elements constitutes a sequence. The present work employs competition to choose an internal representation and time delays to limit the granularity of internal representations. Whether or not a subsequence is represented depends on its frequency of occurrence and the number of units per cluster. As the number of subsequences approaches the number of units per cluster, less frequently occurring subsequences will fail to acquire a unique internal representation.

The effects of self-organization can be further enhanced if the output of one sequence recognition layer provides input to another layer having delays at a different time scale. For example, the motion-detection network described above operates at a very limited time scale with small receptive fields, leading to the linearly independent representation of upward and downward sweeps. Consequently, a network designed to recognize trills would more easily develop trill-selective units given the output of the motion-detection layer than merely sequences of changing tones tones. If the nervous system employs matched spectral and temporal scaling, sequential patterns and feature detectors are probably not analyzed as unitary objects but as constructs arising from several intermediate stages of representation. Individual units exhibit a number of interesting properties, but the value of their response to the system cannot be guaged without understanding the interactions within the sequence recognition layer.

Finally, the simulations demonstrate unit selectivity for sequential patterns: unit response is strongest for complete patterns and becomes weaker as context is removed. Here selectivity is the result of non-optimal bottom-up input enhanced by intracluster competitive dynamics. Beyond this, selective response depends on the elaborate set of lateral connections formed during learning. The tonotopic character of the initial input representation is maintained at all higher layers because units respond weakly to bottom-up activation withing their receptive fields. Of course, a unit selective for a sequence ending on a particular component also responds very weakly to sequences lacking that component, because lateral excitation is insufficient to cause suprathreshold activation. In real neural networks more complex types of learning are probably responsible for unit selectivity. Recent work has shown how the learning of inhibitory connections to short-term memory can enhance unit selectivity. These connections also mediate inibitory response to disordered sequences and superfluous tones like that found in the avian neural response to mal-formed birdsong fragments (Margoliash, 1983).

Altogether, this model demonstrated quite robust performance on complex nonspeech sound patterns – even under severe signal/noise ratios. It may be effective for many real-life acoustic environments.

5 Rhythm: The Adaptive Oscillator Model

Thus far, we have addressed two of the three basic types of temporal information in audition: motion and sequence. In particular, units in the Lexin model developed selectivity to statistically salient frequency-sweeps and sequences. In an open environment, we may expect to see such patterns as tone sweeps that repeat periodically like *tweet, tweet, tweet,...*. Notice that the Lexin model has no way to take advantage of the fact that many important acoustic features, like these *tweets* are periodic or nearly so. In this section, we describe an *adaptive oscillator* model for recognizing "rhythmic" acoustic features. The focus of this section is on the temporal patterning of events, not on their identification. Therefore, we can tentatively assume the representation of an event corresponds only to the onset of that event and its intensity, which we allow to vary between $[0, 1]$ (see Figure 10A). These input patterns are roughly similar to the output of a drum machine with hard and soft beats. Our goal is to construct a mechanism that can tune in to some periodic (or nearly periodic) event and use that predictability to increase its robustness to noise or distracting patterns.

5.1 Description

Three fundamental properties characterize the adaptive oscillator model. **(1)** The adaptive oscillator has a preferred oscillation period which models the listener's preferred IOI (or tempo) (Woodrow, 1951; Fraisse, 1963). **(2)** A fast-acting synchronization procedure models the listener's ability to entrain to salient aspects of a rhythm. Gradient-descent on synchronization error updates the oscillator's preferred period so that it matches periodic components of the input rhythm. **(3)** A decay process acts to oppose synchronization by forcing the adapted oscillator back to its preferred rate. The push-and-pull of decay and synchronization maintain the model's "sense" of the psychological present (Fraisse, 1963). That is, it is easy for the model to adapt to tempos that are around the preferred rate, but increasingly difficult for the model to adapt to tempos that become significantly slower, for example.

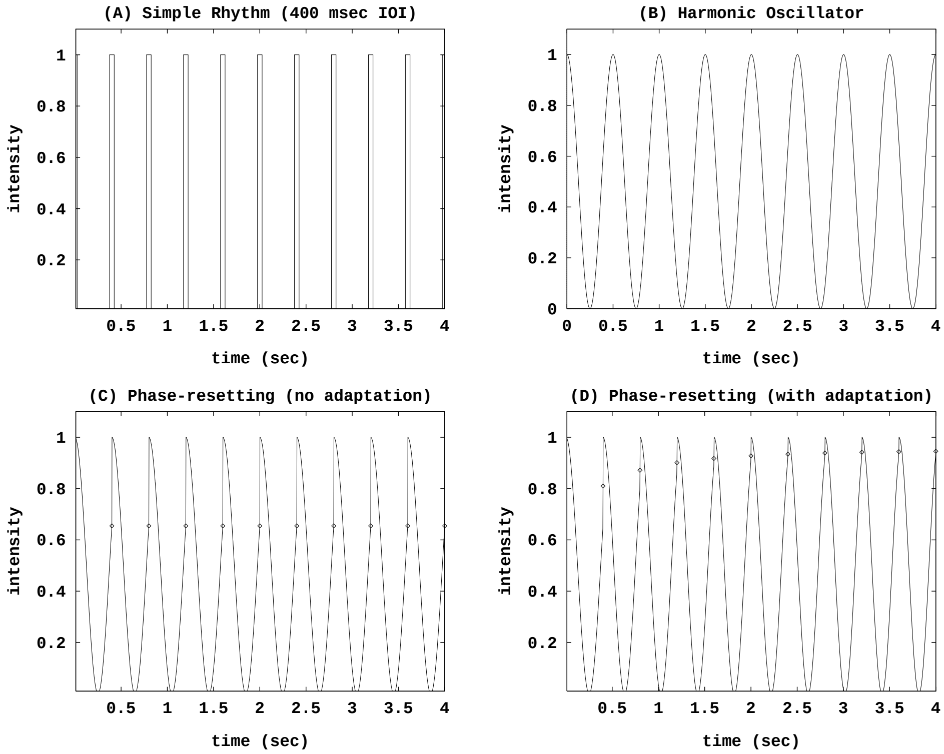

Figure 10: (A) Discrete construction of a simple input rhythm. Each onset occurs every 400 msec and has a duration equal to the sampling period. All other time slices have zero intensity. (B) Eight periods of a 2.0 Hz harmonic oscillator. (C) Input pulses are added to the harmonic oscillator $\phi(t)$ providing a measure of total activity. Total activity exceeds the threshold of 1.0 and forces the oscillator to spike and phase-reset. Output values are marked by a small circle at each phase reset. (D) Fast-acting synchronization is applied to the oscillator. Notice that the output values at each phase-reset continue to increase, providing an entrainment measure.

5.1.1 Preferred Period

For the adaptive oscillator, the preferred period is embodied within a periodic "activation" function, such as the harmonic oscillator

$$\phi(t) = (1 + \cos(\frac{2\pi t}{\Omega(n)}))/2.$$

The period of this oscillator $\Omega(n)$ is initialized to its preferred period: $\Omega(0) = p$ (see Figure 10B). On each time-step, the current input $i(t)$ and the current value of the sinusoid $\phi(t)$ sum together to provide a measure of total activity: $a(t) = \phi(t) + i(t)$ The oscillator generates an output "spike" $o(n)$ each time the total activity $a(t)$ reaches or exceeds a threshold value, $\theta > 0$, after which the activation function is immediately phase-reset to $t = 0$ (see Figure 10C). The threshold value is typically 1.0 and the output $o(n)$ can be simply defined as the value of $\phi(t)$ at that point in time.

5.1.2 Synchronization and Decay

An input pulse that arrives out-of-phase with respect to the zero-phase *spontaneous* firing pattern of the oscillator may, depending on the input intensity, force the oscillator to spike early or late. This phase information can be used to define a spike-driven gradient-descent procedure which synchronizes the spontaneous spiking behavior of the oscillator with rhythmic aspects of the input pattern (see Figure 10D). We define synchronization error as the squared "temporal distance" between input-forced spikes and spontaneous spikes, which is simply the squared difference between the threshold θ and the activation $\phi(t)$:

$$E(n) = 1/2(\theta - \phi(t))^2.$$

To minimize the synchronization error at each output spike $o(n)$ (input-forced or spontaneous), the oscillators period $\Omega(n)$ is adapted by a fraction α that is (negatively) proportional to the partial derivative of the synchronization error $E(n)$ with respect to $\Omega(n)$:

$$\Omega(n+1) = \Omega(n) - \alpha \frac{\delta E(n)}{\delta \Omega(n)}.$$

To include as part of the update rule, a decay process which pushes the adapted period back towards its preferred period p, one assumes that the adapted period $\Omega(n)$ is normally distributed about a mean $\overline{\Omega}$ equal to the preferred period p. By taking the log of this distribution, one obtains a decay term

$$D(n) = -\frac{1}{2\sigma^2}(\Omega(n) - \overline{\Omega})^2$$

which amounts to a penalty for large differences between the adapted period and the preferred period. The standard deviation σ determines how important the decay constraint is. If σ is small, the decay term will be heavily weighted. If it is large, the decay is negligible. The decay term is minimized along with the synchronization error to obtain a modified update procedure which incorporates both the prior constraint that the adapted period decay back to its mean (the preferred period p) and the goal to entrain to the input rhythm:

$$\Omega(n+1) = \Omega(n) - \alpha \frac{\delta E(n)}{\delta \Omega(n)} + \beta \frac{\delta D(n)}{\delta \Omega(n)}.$$

5.2 A General Framework

The above desrcription provides a general framework for investigating adaptive oscillator models. In the simulations, we will investigate four models. Each model is best characterized by its activation function (see Figure 11 and the accompanying Table 1 for a complete desciption of the corresponding equations).

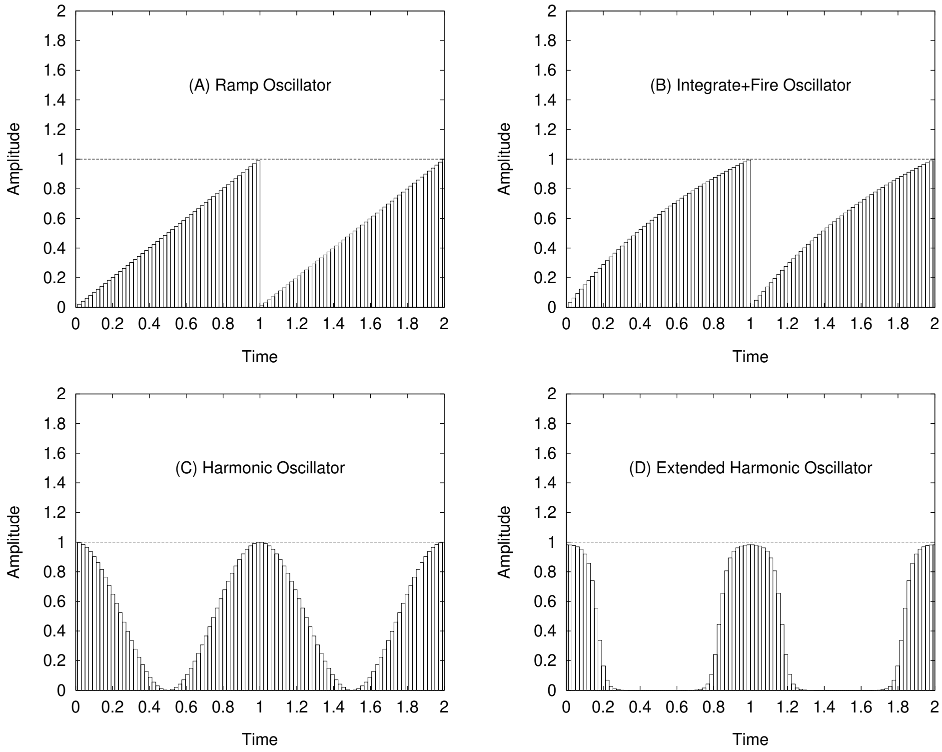

Figure 11: The graphs show activation trajectories for two oscillations of a ramp, integrate+fire, harmonic and extended harmonic oscillator. Each oscillator has a threshold of 1.0 and a preferred oscillation period of 1.0 Hz.

Simulations have evaluated the success of the adaptation procedure (without decay) as it is applied to the four oscillators. The effect of changes in the activation function on oscillator

Oscillator	Activation	Threshold	Pref. Period
Ramp	Ωt	1	$1/\Omega$
Integrate+Fire	$(\Omega+1)(1-e^{-\tau t})$	1	$\ln(\frac{\Omega}{\Omega+1})/-\tau$
Harmonic	$(1+\cos(\frac{2\pi t}{\Omega}))/2$	1	Ω
Ext. Harmonic	$(1+\tanh(\gamma(\cos(\frac{2\pi t}{\Omega})+\beta)))/2$	1	Ω

Table 1: Each model can be categorized as belonging to one of two classes based its activation function shape. (1) The ramp and integrate+fire oscillator have monotonically increasing activation functions with a discontinuity at threshold; i.e., $\phi(t)$ at phase-reset changes from 1.0 to 0.0. (2) the harmonic and extended-harmonic oscillators have continuous non-monotonic activation functions.

dynamics are discussed in the next section, in the context of developing a bank of oscillators which develop sensitivity to the underlying meter of simple rhythmic patterns.

Although decay is not included as part of these simulations, in recent work (McAuley, 1994b) we have shown that decay plays an important role in modeling certain results on human time perception. Specifically, we compared the dynamics of an adaptive harmonic oscillator with decay to human data on tempo discrimination. The model was trained using a simulated 2AFC adaptive-tracking procedure to discriminate tempo changes across a range of inter-onset intervals (100-1500 msec). It has been shown that human subjects, performing the task described above, exhibit maximum tempo sensitivity–as a function of the number of isochronous intervals–between 300-800 msec (Drake and Botte, 1993). The performance of the adaptive oscillator model was shown by (McAuley, 1994b) to provide a qualitative match to the Drake and Botte data.

Notice that unlike a forced oscillator, the period of an adaptive oscillator is incrementally modified so that after cessation of stimulation, oscillation continues at a rate that is entrained to a frequency component of the input. With decay, the oscillation period will gradually return to the oscillators preferred period. Thus, adaptive oscillators embody a temporal expectation for a continuing pattern of input pulses, which is similar to the notion of "pulses of attention" described by (Jones, 1976).

5.3 Simulation 1: Adaptive Oscillator Dynamics

The first simulation explored how well each oscillator entrained to input patterns with a constant drum-pulse amplitude (see Figure 10A). To be able to compare across models, all shared parameters were fixed at the same values. Each oscillator's period was initialized to 0.5 seconds and the input pulse amplitude was fixed at 0.5. The responses of each oscillator to input patterns with periods ranging from 0.01 to 0.99 seconds in steps of 0.01 were recorded after the learning algorithm converged, or after it was clear that the algorithm would not converge. Convergence time varied over a wide range (two to one hundred 'model' seconds). Within a particular model, the main variable was the learning rate, with the largest learning rates yielding the fastest convergence times. Not surprisingly, large learning rates also tended to produce very erratic behavior by jumping the oscillator spikes back and forth over the entraining input pulses. Small learning rates could still produce non-convergent "chaotic" behavior, in which case the data point was thrown out, so as to only show the input periods for which each oscillator entrained its response.

The results for a learning rate of 0.05 are shown in Figure 12. These graphs share a couple of very noticeable properties. An oscillator's response to a sweep of input periods often falls along a straight line. These line segments correspond to entrainment regions, with the entrainment ratio $m : n$ (number of input pulses per oscillator pulse specified by the slope of the segment. 'Breaks' between line segments are bifurcations in the oscillators response. The shape of the oscillators activation function is one of the main variables that determines the size and location of each entrainment region. This is discussed in some detail below, for each of the investigated activation functions, in the context of developing a model for rhythm perception.

In developing a model which captures aspects of human rhythm perception, the obvious question is what *should* the entrainment regions look like. Unfortunately, this does not afford an easy answer, but the desired form of the oscillators adaptive response should have certain features. In the context of a larger system responding to the underlying meter of a rhythmic pattern, it is desirable for the oscillators to have entrainment ratios other than $1 : 1$ (corresponding to multiple line segments in the Figure 12); e.g., a bank of oscillators

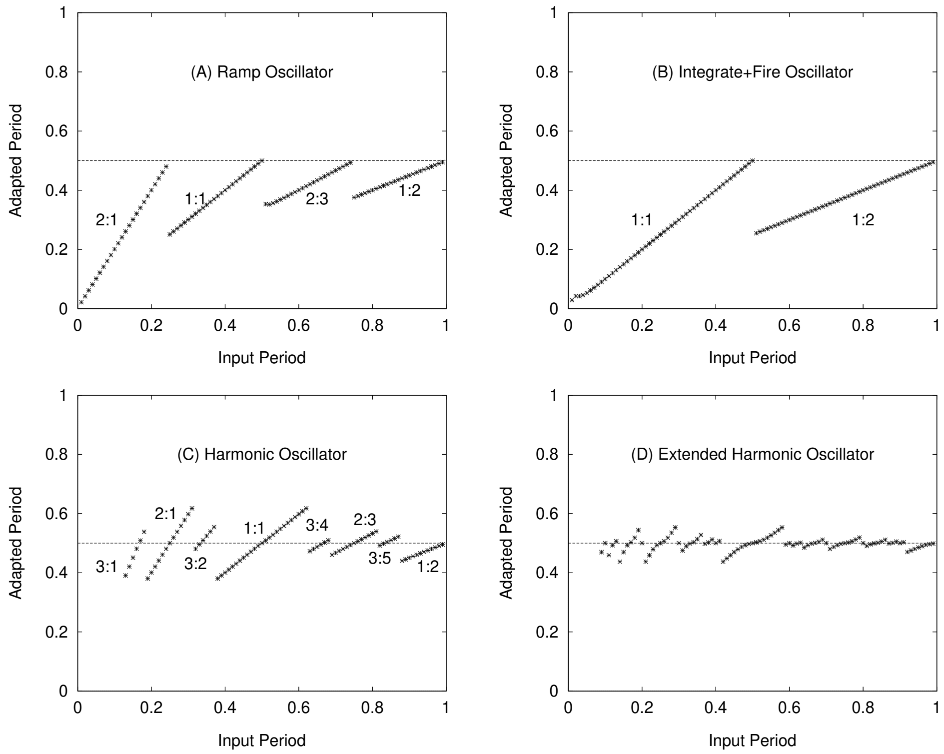

Figure 12: These graphs compare the entrainment ratios and bifurcation points of the ramp, integrate+fire, harmonic, and extended harmonic oscillators for 0.5 amplitude input patterns. Along each x-axis is a 'sweep' of input periods ranging from 0.01 to 0.99 in steps of 0.01. Along each y-axis is the adapted period of the oscillator. The resting period of each oscillator is initially 0.5 seconds and the learning rate is fixed at 0.05.

responding to a 'strong-weak' pattern should adopt $1 : 1$ and $2 : 1$ entrainment ratios, at the beat and measure level respectively[5]. Since our perception of meter is not affected by modest variations in tempo, a viable oscillator model should also be able to track tempo changes by being able to speed up and slow down over a modest range without encountering a bifurcation. In the graphs, this corresponds to 'medium' length line segments which extend above and below the resting period of 0.5 seconds.

The ramp and integrate+fire oscillators are unsuitable models because they can only adapt to an input pattern by speeding up. The derivative of their activation functions is always positive resulting in a $\Delta\Omega$ that always has the same sign. If the input pattern

[5]Beat and measure refer to perceived segmentation of time and grouping in time, respectively. The timing of notes in a musical performance is based on the perception of an underlying beat duration. How these notes are grouped together in time is a function of the beat duration and refers to the measure length.

slows down, these models have to find a different entrainment ratio. Notice also that the integrate+fire oscillator shows fewer bifurcations. This occurs because its activation function increases at a faster rate than the constant slope ramp oscillator. The net result is an 'effective' increase in input amplitude. Although it is not shown in these graphs, an actual increase in input amplitude will decrease the number of bifurcations. A simulation was run in which the input amplitude was increased to 0.8. The entrainment ratios of the ramp oscillator for this condition matched the entrainment ratios for the integrate+fire oscillator depicted in Figure 12B.

Unlike the ramp and integrate+fire oscillators, but more like human listeners, the harmonic and extended harmonic oscillators can adapt to tempo changes by speeding up and slowing down their responses. For these models, this implies that $\Delta\Omega$ can be both positive and negative. Notice however, that a large portion of the extended harmonic's activation function is flat (see Figure 11D). Since the error gradient for flat curves is almost zero, this oscillator does not adapt to input-forced spikes which occur during the flat segment. If instead, an input-forced spike occurs during the steep portion of the activation function, the system adapts very quickly. The main consequence of having very flat and steep activation function segments is to implement a form of *temporal receptive field*. That is, the adapted period of the extended harmonic oscillator tends to remain within a small 'window' of its initial resting period (see Figure 12D). As a result, it must find more exotic entrainment ratios to minimize error. This greatly increases the number of bifurcation points and makes tracking a pattern which changes tempo very difficult.

On the other hand, the activation function for the harmonic oscillator is not as flat. For the most part, its slope is roughly a constant positive or negative non-zero value. Consequently, the harmonic adaptive oscillator has fewer bifurcations, is more stable, and converges fairly quickly. It is a more plausible model for tracking tempo changes than the extended harmonic oscillator model. In the next section, we explore the ability of the harmonic oscillator model to track tempo changes and maintain sensitivity to metrical structure in a way that is similar to humans.

5.4 Simulation 2: Meter and Expressive Timing

It is the purpose of this simulation to show that a cluster of four adaptive harmonic oscillators with a range of initial periods can adapt to the beat and measure level components of two simple rhythmic patterns and that the entrainment ratios these oscillators adopt is sensitive to the meter of the stimulating rhythm. Furthermore, when the tempo of the rhythm drifts, the oscillators are able to track the change without losing their 'sense of meter'. Such expressive timing deviations are common to musical performance (?).

The oscillators received only unweighted bottom-up input; lateral connections between oscillators were not included. In a more complete system, oscillators would have lateral connectivity and would develop sensitivity to metrical structure as the result of two interactive learning procedures: a gradient descent entrainment process (as discussed in the previous section) and a Hebbian learning process for adjusting lateral connection strengths (Hebb, 1949). By definition, Hebbian learning would reinforce the connections between the oscillators entrained to the beat-level pattern and the measure-level pattern each time the ensemble 'fires' in-phase (once every two beats for a duple pattern). Note that by virtue of their initial resting periods, some oscillators in the group will not entrain to either the beat or measure level, but may in fact discover a more exotic entrainment ratio. This is problematic if these oscillations conflict with a description of the meter. There are at least two potential ways of suppressing the unwanted oscillations. Connections to oscillators not responding the beat or measure level could be weakened by virtue of the fact that their entrainment patterns do not correlate well with the input amplitude pattern. A second approach would be to use the effect of a substantial number of oscillators responding to the beat and measure level to 'swamp' the extraneous oscillations. The design and implementation of lateral connections using this second approach is in progress.

The two rhythms in this simulation were a strong-weak 'duple' pattern with amplitudes alternating between 0.5 and 0.2 and a strong-weak-weak 'triple' pattern described by the repeating amplitude sequence 0.5, 0.2, and 0.2. The initial tempo for these patterns was 143 beats per minute which corresponds to a beat period of approximately 0.42 seconds. To change the tempo of the rhythms, a random number of zero amplitude time-slices were added to the beat period of each measure. This procedure was cumulative so that the new tempo of each measure varied from the previous measure's tempo by at most 10%. The net result was a gradual drift in tempo. The tempo changes were the same for both rhythms so as to compare the responses of the oscillators. For the simulation, the tempo varied for 150 measures. This pattern of tempos is shown in Figure 13A and B as a solid line by plotting the beat period of the input as a function of the measure number.

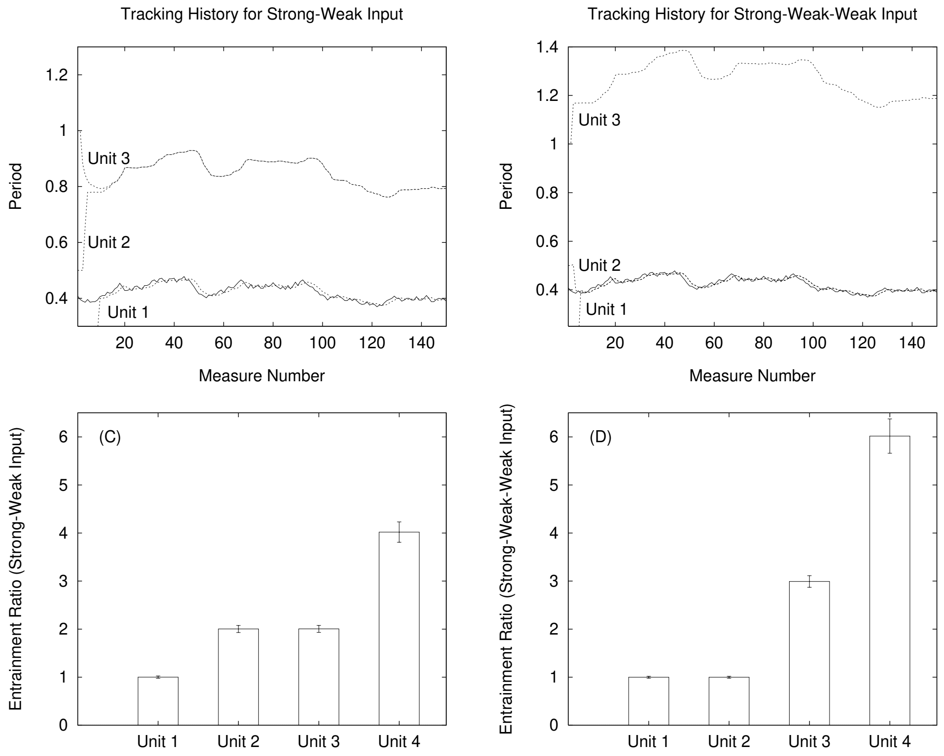

Figure 13: (A) and (B) show the ability of a bank of four harmonic oscillator units to track tempo changes to two simple rhythmic patterns. (C) and (D) show that the entrainment ratio profiles in response to these rhythms is sensitive to the underlying meter. This property is independent of tempo.

The periods of the four oscillators were initialized to 2.0, 1.0, 0.5, and 0.25 seconds; in terms of musical tempo, this corresponds to 30, 60, 120, and 240 beats per minute respectively, ranging from very slow to very fast. These choices were somewhat arbitrary and were not selected so as to model specific psychophysical data.

So that the oscillators would adapt quickly, the learning rate was set to 0.4. The dotted lines in Figure 13A and B, labelled Unit 1, Unit 2, and Unit 3, show the tracking histories for the first three oscillators in response to the 'duple' and 'triple' rhythms respectively. For both rhythms, all of the oscillators (including Unit 4 which is not shown) converged to stable entrainment ratios after only a few measures, and were able to track subsequent shifts in tempo. This is manifest by comparing the tracking history trajectories and the drift pattern of the tempo.

With respect to the specific entrainment ratios, it is important to make two observations which are reflected in Figure 13C and D. These bar-graphs compare the entrainment 'profiles' of the bank of oscillators in response to the duple and triple rhythms. Each oscillator is labelled as a unit number along the x-axis. Bar height indicates the corresponding entrainment ratio. Associated with each ratio is an error-bar indicating the by-measure variance in the entrainment ratio as the unit adapted to changes in tempo.

The first observation is that a unit-by-unit comparison between the entrainment profiles reveals some important similarities and differences. Unit 1 maintains a $1:1$ response to both rhythms, providing a beat-level description of the underlying meter of the input rhythm. On the other hand, Units 2, 3, and 4 switch their responses to each rhythm, but these different entrainment ratios are consistent with an appropriate metrical description of the input rhythm. In particular, Unit 3 entrains $2:1$ to the strong-weak rhythm, but it entrains $3:1$ to the strong-weak-weak pattern, providing a measure-level description of the underlying meter of both rhythms. The second observation is that the error-bars provide a quantitative measure of how well each unit is able to adapt to changes in tempo. The small error-bars indicate that in spite of modest timing deviations, the units collectively maintain a consistent description of the meter.

5.5 Conclusions

We have described a general class of *adaptive oscillator* models for processing rhythmic patterns. All models in this class have the same basic building blocks and use gradient-descent to entrain to periodic input patterns. Of the four models investigated in the simulations, the adaptive harmonic oscillator is best suited for modeling aspects of human rhythm perception.

In particular, a version of the adaptive harmonic oscillator model, that incorporates decay to preferred frequency, provides a qualitative match to data on tempo discrimination. This model is also capable of maintaining sensitivity to the meter of simple rhythmic patterns, despite modest changes to the pattern tempo. It is likely that such 'hardware' designed specifically to encode timing relationships is likely to be more useful for rhythm recognition that generic recurrent neural networks.

At present, the adaptive oscillator approach to rhythm is in the early stages of development. There is a clear need for formalizing the mathematical properties of these models. The description of a general class of adaptive oscillators is a step in this direction. As a next step, a careful comparison needs to be made between the adaptive oscillator paradigm and earlier work which investigated the dynamics of forced oscillation (Keener et al., 1981; Stiber, 1992). This would establish a better foundation for designing models with specific constraints on entrainment ratios and bifurcations, as suggested by the myriad of psychological data on rhythm and timing.

6 Toward an Integrated Auditory Processing Model

We have discussed thus far a number of considerations that are relevant to development of an auditory system that can adapt to a relatively open auditory environment. First, we considered the practical question of the kind of information that is likely to be useful. Then we discussed some advantages and drawbacks of a number of previous approaches and proposed that a useful system would have to include, in addition to a static featural description, methods for recognizing acoustic motions and rhythmic structures as well as sequences in order to do a satisfactory job at auditory perception. In this section, we will describe two things. First, the Auditory Environment Generator we are developing at IU – a tool for testing the capabilities of an auditory system, and second, some ways in which the two models described above might be integrated into a General Auditory System that is capable of recognising complex hierarchical periodic and nonperiodic patterns in sound.

Auditory Environment Synthesizer with Realtime Audio Input. Although training any network may be very time consuming on a computer, the actual performance of these models in processing spectral inputs can easily run in realtime on a laboratory workstation. Thus it is quite feasible to outfit a workstation to respond using realtime audio input – such as hand claps, speech or even the music from the radio.

Realtime interaction by the researcher with the model should allow important insights into the behavior of the model, and would permit direct and vivid evaluation of its robustness.

We are outfitting a workstation with an auditory input system that will simulate basic properties of the ear and output a seemingly continuous (though, of course, sampled) signal. Signal processing software will filter the signal through 10-20 frequency channels, retaining the temporal aspects of auditory signals that are apparent in the auditory nerve (Shamma, 985b; ?). This means that the software will be fed a spectrum-like signal sampled 10 to 20 times per second. Models under investigation will be fed such input from the auditory environment programs or from a microphone in the lab. This system will serve as a testbed for all the various network models to be tested.

The Acoustic Environment Synthesizer will be based on *C-speech* programs that produce a variety of beeps, glissandos, noise bursts at various center frequencies. Each event will have some probability of occurrence that is independent of each other event. Events can occur periodically as well. Any given model can be tested in a range of environments and can be allowed to adapt to each environment for varying amounts of time.

These two systems will permit vivid testing of any model and its degree of robustness.

Integrated Perceptual Model. The next figure shows some initial ideas for combining the Lexin model for recognition of familiar sequences with the Adaptive Oscillator model. Basically, the model implements ideas of Mari Jones about the nature of periodic attention. On the basis of research on patterns that do or do not appear at 0 phase in a periodic pattern, Jones concludes that periodic patterns induce attention that is 'stronger' at certain phase angles than at others – that is, that subjects exhibit greater sensitivity to events at 0 phase than to events elsewhere in time.

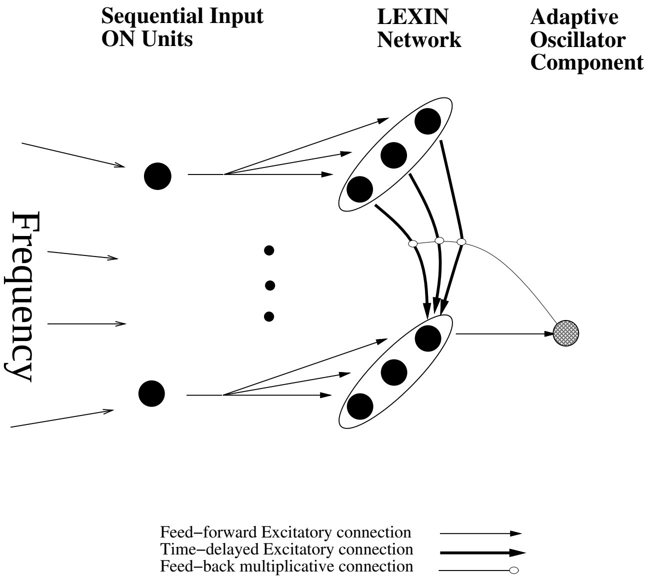

Figure 14: Proposed integration of adaptive oscillators with Lexin. Each feature-detecting node will have an adaptive oscillator that can gate its inputs to the feature detector at times determined by the current firing rate of the oscillator.

The single module in Figure 14 shows an adaptive oscillator attached to each node of the competing cluster of features at each frequency. Each of these nodes responds to a sequential fragment. If one node in the cluster is activated, it tries to excite its adaptive oscillator. If the oscillator is excited periodically, it becomes more active, producing increased output at its current firing frequency. And of course, if the current period is off somewhat, it begins to adapt to that period. Within a few cycles, the increased activity of the oscillator should lead to increase in gain to all excitatory leads into that node, thus lowering the firing threshold for that node. This means that after the node that has been excited several times from bottom up, it will now find it easier to fire – even in serious noise or during competing frequency patterns. Thus this model should exhibit a reduced threshold for its input pattern once the resonator has been able to adapt.

Eventually, one might want to do more complete integration of these 'functional modules'. In any case, it is clear that with a suitable test bed, the general efficacy of such a model can be tested.

7 Conclusions

We have tried to clarify how the problem of auditory temporal patterns faced by animals is different in fundamental ways from the most general problem of so-called sequence analysis. In particular, a nervous system can afford to concentrate primarily on familiar patterns. Furthermore, we suggested that familiar sound patterns can be classified into three basic types of information: acoustic motions over an infinitessimal time interval, periodic or rhythmic patterns and plain old sequential patterns. Complex structures like music and speech can be viewed as constructed out of these basic types. We presented the Lexin model that self-organizes to recognise common sequential patterns in its auditory environment. Further development of models of this class combined with adaptive rhythm recognisers should permit rather sophisticated inerpretation of a wide range of acoustic environments for which the basic architecutre was in no way specialized.

References

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

Amari, S. (1972). Characteristics of random nets of analog neuron-like elements. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2:643–657.

Anderson, S. (1992). Self-organization of auditory motion detectors. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 684–689, Hillsdale, NJ. Lawrence Erlbaum Assoc.

Anderson, S. (1994). *A Computational Model of Auditory Pattern Recognition.* PhD thesis, Indiana University.

Anderson, S., Merrill, J., and Port, R. (1988). Dynamic speech categorization with recurrent networks. In *Proceedings 1988 Connectionist Summer School*, pages 398–406. Morgan-Kaufmann, San Mateo, California.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Bradford Books, MIT Press.

Carpenter, G. and Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 37:54–115.

Chappell, G. and Taylor, J. (1993). The temporal Kohonen map. *Neural Networks*, 6:441–445.

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.

Cummins, F., Port, R., McAuley, J., and Anderson, S. (1993). Temporal pattern recognition with fully-recurrent networks. In *Research in Auditory Cognition*, pages 27–40, Bloomington, Indiana. Cognitive Science Program, Indiana University. No. 6.

Dehaene, S., Changeux, J., and Nadal, J. (1987). Neural networks that learn temporal sequences by selection. *Proc. Natl. Acad. Sci., USA*, 84:2727.

Drake, C. and Botte, M.-C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception and Psychophysics*, 54(3):277–286.

Elman, J. (1989). Structured representations and connectionist models. In *Proceedings of the 11th Annual Conference of the Cognitive Science Spciety*, pages 17–25, Hillsdale, NJ. Lawrence Erlbaum Associates.

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

Elman, J. and Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, 83:1615–26.

Fraisse, P. (1963). *The psychology of time.* Lowe and Brydone, London.

Gottlieb, G. (1975). Development of species identification in ducklings: II. Experiential prevention of perceptual deficit caused by embryonic auditory deprivation. *Journal of Comparative and Physiological Psychology*, 89:675–684.

Green, D. (1976). *An Introduction to Hearing.* L. Erlbaum.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. *Biol. Cybernetics*, 23:121–134.

Hebb, D. O. (1949). *The Organization of Behaviour.* John Wiley and Sons, New York.

Hertz, J., Krogh, A., and Palmer, R., editors (1991). *Introduction to the Theory of Neural Computation*, volume I of *Sante Fe Institute Studies in the Sciences of Complexity.* Addison Wesley, New York.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci., USA*, 79:2554–2558.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, 83:323–355.

Jordan, M. (1986). Serial order. Technical Report 8604, Institute for Cognitive Science, U. of California at SanDiego, La Jolla, CA.

Keener, J. P., Hoppensteadt, F. C., and Rinzel, J. (1981). Integrate-and-fire models of nerve membrane response to oscillatory input. *SIAM Journal of Applied Math*, 41(3):503–517.

Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proc. Natl. Acad. Sci., USA*, 83:9469–9473.

Kleinfeld, D. and Sompolinsky, H. (1988). Associative neural network model for the generation of temporal patterns. *Biophysical Journal*, 54:1039–1051.

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

Kurogi, S. (1987). A model of neural network for spatiotemporal pattern recognition. *Biol. Cybernetics*, 57:103–114.

Kurogi, S. (1991). Speech recognition by an artifical neural network using findings on the afferent auditory system. *Biol. Cybernetics*, 64:243–249.

Lang, K. J., Waibel, A. H., and Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:23–43.

Leek, M. R. and Watson, C. (1984). Learning to detect auditory pattern components. *Journal of the Acoustical Society of America*, 76:1037–1044.

Margoliash, D. (1983). Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J. Neurosci.*, 3:1039–1057.

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

McAuley, J. D. (1994a). Finding metrical structure in time. In Mozer, M. C., Smolensky, P., Touretzky, D. S., Elman, J. L., and Weigend, A. S., editors, *Proceedings of the 1993 Connectionist Models Summer School*, pages 219–227. Lawrence Erlbaum.

McAuley, J. D. (1994b). Time as phase: A dynamic model of time perception. Unpublished manuscript.

McClelland, J. and Elman, J. (1986). Interactive processes in speech perception: The trace model. In McClelland, J. and Rumelhart, D., editors, *Parallel Distributed Processing*, volume 2, chapter 15, pages 58–121. MIT Press, Cambridge.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Moore, D. R., Hutchings, M. E., King, A. J., and Kowalchuk, N. E. (1989). Auditory brain stem of the ferret: Some effects of rearing with a unilateral ear plug on the cochlea, cochlear nucleus, and projections to the inferior colliculus. *J. Neurosci.*, 9:1213–1222.

Mozer, M. C. (1993). Neural net architectures for temporal sequence processing. In Weigend, A. and Gershenfeld, N., editors, *Predicting the Future and Understaning the Past*, Redwood City, California. Addison-Wesley.

Port, R. (1986). Invariance in phonetics. In Perkell, J. and Klatt, D., editors, *Invariance and Variability in Speech Processes*. Erlbaum Associates, Hillsdale, New Jersey.

Port, R. and Anderson, S. (1989). Recognition of melody fragments in continuously performed music. In Olson, G. and Smith, E., editors, *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society*, pages 820–827, Hillsdale, NJ. L. Erlbaum Assoc.

Port, R. and Cummins, F. (1992). The English voicing contrast as velocity perturbation. In *Proceedings of the Conference on Spoken Language Processing*, Edmonton, Alberta. University of Alberta, Personal Publishers.

Port, R. and Dalby, J. (1982). C/V ratio as a cue for voicing in English. *Journal of the Acoustical Society of America*, 69:262–74.

Port, R. F. (1990). Representation and recognition of temporal patterns. *Connection Science*, 2.

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

Robinson, T. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5:259–274.

Schreiner, C. and Langner, G. (1988). Coding of temporal patterns in the central auditory nervous system. In Edelman, G., Gall, W., and Cowan, W., editors, *Auditory Function: Neurological Bases of Hearing*, pages 337–362. John Wiley and Sons, New York.

Sejnowski, T. and Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168.

Shamma, S. (1987). Neural networks for speech processing and recognition. In *Proceedings of the International Conferece on Neural Networks*, pages IV:397–405. IEEE.

Shamma, S. A. (1985b). Speech processing in the auditory system ii. lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, 78(5):1622–1631.

Smythe, E. J. (1988). Temporal computation in connectionist models. Technical Report 251, Indiana University, Computer Science Department, Indiana University, Bloomington, Indiana.

Sompolinsky, H. and Kanter, I. (1986). Temporal association in asymmetric neural networks. *Physical Review Letters*, 57(22):2861–2864.

Stiber, M. (1992). *The Dynamics of Synaptic Integration*. PhD thesis, UCLA.

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

Tank, D. and Hopfield, J. (1987). Neural computation by concentrating information in time. *Proc. Natl. Acad. Sci., USA*, pages 1896–1900.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1987). Phoneme recognition using time-delay neural networks. Technical Report TR-1-006, ATR Interpreting Telephony Research Laboratories.

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

Watrous, R., Ladendorf, B., and Kuhn, G. (1990). Complete gradient optimization of a recurrent network applied to /b/, /d/, /g/ discrimination. *Journal of the Acoustical Society of America*, 87:1301–1309.

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

Woodrow, H. (1951). Time Perception. In Stevens, S. S., editor, *Handbook of Experimental Psychology*, pages 1224–1236. Wiley, New York.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. *Neural Comp.*, 3:179–193.