12 Naive Time, Temporal Patterns, and Human Audition

Robert F. Port, Fred Cummins, and J. Devin McAuley

EDITORS' INTRODUCTION

Change over time is, in many ways, the raw material of perception. In no modality is this more obvious than in audition. Much if not most of the information that is contained in auditory events like speech is a matter of the way that a signal changes over time. Humans and other animals have remarkable abilities to extract this information from the auditory signal in real time, abilities which far exceed anything available in current technology. A key problem for cognitive scientists is to figure out how natural cognitive systems do it.

In thinking about processes that unfold in time, we are accustomed to applying an objective or absolute measure like the second. A clock marks the passing of seconds, and for a process to happen in time is for the events that make it up to be laid out against this independent yardstick. This way of conceptualizing processes in time is so obvious that it is difficult to see what the alternative might be. Yet Port, Cummins, and McAuley begin this chapter by arguing that this standard approach, which they dub the "naive" view of time, is not particularly useful if our aim is to understand how natural cognitive systems perceive auditory patterns. The use of natural time as a standard in perceptual models typically requires positing a buffer for a raw stimulus trace, in which unit intervals of time are transformed into units of space. The authors argue strongly against the possibility of any such buffer in the auditory system. Furthermore, absolute measurements would not be the most useful basis for recognizing temporal events anyway. Somehow, our auditory systems manage to handle information contained in processes that unfold over time without reliance on buffered sensory traces or on measurements in absolute (millisecond) units.

Port, Cummins, and McAuley argue that temporal information comes in two basic varieties: serial order and durational information. The first is a familiar feature of language: one thing we want to do when hearing what another says is extract from the auditory signal the words or phonemes in the order that they arrive. Most standard computational models for automatic speech recognition, such as hidden Markov models, attempt to obtain this serial order information by abstracting across as much irrelevant durational and rate variation as possible. In the process, however, they typically run into a severe problem of state proliferation in the model. This is because each possible variant must, in effect, be spelled out explicitly in the model. In this chapter the authors present a simple dynamical model for serial order extraction which is able to avoid this difficulty.

For other purposes, nuances of duration are very important. How can a natural cognitive system pick up on the rhythm or period of an auditory signal without prior knowledge of either signal identity or signal rate? In the latter part of their chapter, Port, Cummins, and McAuley present another novel dynamical model, an oscillator which automatically latches onto the period of a stimulus signal, even if this period is irregular or noisy. They propose that the period this oscillator extracts, which is a pattern intrinsic to the signal itself, be used as the standard against which relative duration measurements pertinent to recognizing an input pattern are made. This is practical, since listeners need a standard from somewhere, and the auditory signal to be recognized is always an available source, and it is desirable, because the signal's own period is a more useful measure than the second for patterns like speech, animal gaits, music, and so forth, which can occur at a range of rates.

12.1 INTRODUCTION

This chapter is about time and patterns in time. We are concerned with patterns that can be defined only over time and may have temporal constraints as part of their definition. Although there is a widespread view that time can be treated by nervous systems in the same way it is treated by scientists and engineers, we argue that this approach is naive—that there is no *general* method for representing time in the nervous system, i.e., no single representational mechanism that is applicable to recognition of all temporal patterns. Instead, different patterns are analyzed using different methods of measuring temporal extent. We then present several dynamic mechanisms developed in our laboratory for the recognition of various kinds of patterns in time.

12.2 TIME AND TEMPORAL PATTERNS

Time is one of the slipperiest of concepts to talk about. Everything takes place in time, from the history of the planet to the movements of our body; even our various attempts to talk (or think) about time happen in time. It is often said that the world has "things" and "events." The things endure through time without changing much. Events occupy some amount of it, whether fleeting or glacial. Despite the inexorability and continuity of time, we seem nevertheless to have a contrary intuition that there is a "now," a region in time surrounding the present where things are not changing—where most events stand still for us. This is where the various sciences want to live—where everything can be described now and yet we can have some confidence that the description will hold for the future as well. Of course, the notion of now, as a static description of events, is always understood as assuming some particular time scale, from *seconds* to *years*. We know that anything static can be seen to involve change if it is looked at over a longer time scale. Conversely, most static things also turn out to have a temporal component when examined on a shorter-than-usual time scale. Thus, solid material objects, color and words, for example, all have temporal structure if looked at either on a very short or very long time scale.

Things that seem intuitively to *happen in time* are events that last longer than a quarter of a second or so. We will call this time scale the "cognitive time scale." It is the time scale over which humans can act; the scale at which events are slow enough that we might grab with our fingers or blink an eye. The timing of events shorter than this often plays a major role in perception, though we are typically not aware of the role of temporal detail in their specification. Thus, if color recognition depends on the frequency of certain waves of energy, color does not thereby become a temporal pattern at the *cognitive* time scale. Typically, we can observe the temporal properties of very short (subcognitive) events only with special technology. Time has certain obvious similarities to physical distance, such as continuity. We talk of events being "near" or "far" in the past or future just as naturally as we use such terms for physical distance. Like physical distance, we can control our own physical activity down to a certain duration: eye blinks and experimental reaction times (around a quarter of a second) are about as fast as we can move. Of course, unlike physical distance, completed events that are now far away can never become near again. Science fiction fantasies like time travel base their intriguing pseudoplausibility on the trick of taking time as if it were *really* reversible—just like positions along a line drawn on the ground: move forward and then move back.

Before discussing these issues any further, it will fix ideas if we specify a few concrete examples of cognitive auditory patterns, the domain we address in this paper. These patterns can be defined only over time and their temporal extent is normally perceived as a time-extended event by humans. These are the kind of auditory events for which we seek plausible recognition mechanisms. Since we are concerned about recognition, it is also critical to clarify what kind of variations in each pattern are irrelevant to pattern identity and what kinds may cause reinterpretation.

1. Consider the sound of a large quadruped locomoting in some gait or other. The trot of a horse sounds quite distinct from a walk or gallop. Since each gait can occur over some range of rates, simply measuring the time periods between footfalls in *milliseconds* will not by itself allow a representation of a gait that will be invariant across different rates. Clearly, the characteristics of trot or gallop must be specified in terms that are relative to the other events in the sound pattern itself.

2. The theme of *The Merry Widow Waltz* is a temporal pattern defined by a particular melody (i.e., by a sequence of tones from the Western musical scale) played with a particular rhythmic pattern that fits in the waltz meter of three beats per measure. This pattern would be "the same tune" even if played in a different key or if played somewhat faster or slower. On the other hand, if we increased its rate by a factor of 4 or more, or if we severely modified the rhythm (by changing it to a 4/4 meter, for example), we would find that the identity of the tune was destroyed.

3. The spoken word "Indiana" normally has a stress on the [æ] vowel. The word is still the same even when spoken by different voices at a range of speaking rates (up to a factor of about 2 faster or slower). In fact, one could change the stress and say "IN-diana," stretching or compressing various internal portions of the word in time by 15% to 20% and still have it be easily recognizable. Eventually, of course, by such a process one would do severe damage to its linguistic identity for a speaker of English (Tajima, Port, and Dalby, 1993).

4. The spoken sentence "I love you" is also a temporal pattern, although a sentence allows much wider leeway than words or melodies in the actual layout of the events in time. Temporal detail plays a much smaller role in the specification of the structure of a sentence than it does for tunes or spoken words. It seems that they just need to appear in a certain serial order. Still, temporal details are known to affect the parsing listeners construct for ambiguous utterances. Thus, if you read aloud $2(3^2)$, it will be quite different from $(2 * 3)^2$. The difference between them is best described in terms of the location of valleys and peaks in the instantaneous speaking rate, brief decelerations or accelerations that lengthen or shorten speech segments along with any silence. It is not usually a matter of pauses or silent gaps, as it is often described.

It is clear that each of these examples can be defined only over time. Thus a recognition mechanism must collect information over time so that decisions can be delayed long enough to be meaningful. Each of these patterns has a different set of temporal constraints on the essence of the pattern. Still, for all of these examples, the pattern can occur at different rates even though the rate change is of low importance in comparison with the importance of the durational relations internal to each pattern. What kind of mechanism enables humans to recognize such patterns? The view we defend is that nervous systems adopt a variety of ad hoc strategies for describing the temporal structure of patterns.

Living in Time

What is the relevance of time to animals like us? It is critical to differentiate two major uses of the word *time*. First there is history, the Big Pointer, we might say, that persistently moves our lives forward. And then there is time as information about the world. In the latter role, events in time happening now must be related to "similar" events that occurred earlier—either to the individual or to the species. An animal needs to be able to find certain patterns and structure in events that occur in time. To do this requires neural mechanisms for recognizing when an event recurs. Many kinds of clocks, for example, have been found in animals and plants that track the cycles of the sun, both through the day and through the year. Modern science has developed an absolute time scale for temporal description; a scale that depends on the notion of historical time moving forward incessantly. One question is whether animal nervous systems also have accurate mechanisms of this sort.

Scientific Time In order to address questions about the physical world over the past few centuries, Western science has developed various mechanical and mathematical tools for measuring and describing time as an absolute variable. Given some form of clock, scientists and other "moderns" can treat time as just another dimension, one that resembles one of the three dimensions of Euclidean space. Instead of meters, we agree on standard units (second, day, year) to provide a basis for absolute measurement. Mathematically, one seldom needs to treat $f(t)$ as different in any way from $f(x)$. We do not hesitate to plot time on the x-axis of a graph displaying temperature, air pressure, velocity, or any other quantity that is measurable over a small Δt. From such displays of waveforms and spectra, we are able to study the properties of many kinds of events: economic cycles, cardiac signals, the motion of physical objects, sound waves, etc. For example, figure 12.1 shows a sound spectrogram of the phrase "mind as motion" spoken by an adult male. Time is one axis and frequency the other. The darkness of stippling shows the amount of energy in rectangular $\Delta f \times \Delta t$ cells of size 300 Hz × 3 ms. Such displays have become almost second nature to us, and have become integral components of modern thought. Most Americans these days are quite comfortable with stock reports and monthly rainfall graphs. One empirical question that arises for biology and cognitive science is this: *To what extent do the auditory systems of animals employ a display of energy by time in support of the recognition of sound patterns at the cognitive time scale?*

At very short time scales (under a millisecond) measures of time in absolute units like microseconds play a major role in measuring the direction of sound sources using time lags between the two ears (Shamma, 1989). This is, of course, a "subcognitive" phenomenon. But what about at the longer, cognitive time scale; the time scale suitable for recognizing events like words? We propose that in an auditory system for cognitive processing, time is not treated as just another spatial dimension. Spatial axes, like time in a plot of temperature, are reversible. Unlike actual time, one can scan the graph in either direction. If you were using a template to look for a pattern in such a display, you could simply slide the template back and forth until an optimal match were found. But such scannable displays are human artifacts. They are generated by an "assignment clock," some device that moves a sheet of paper past a pen point at a constant rate, or that places equally spaced time samples in a computer buffer, thereby creating a record of "instantaneous" values of the parameter (or, more accurately, values averaged over a short time interval). Since motion across the paper or placement in the buffer is done at a constant rate, distance along the buffer serves as a reliable measure of absolute time.

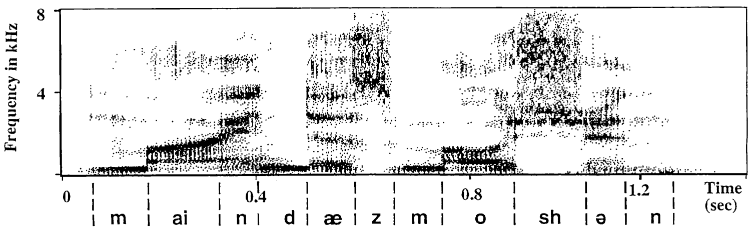

Figure 12.1 A sound spectrogram of the utterance "mind as motion," where the x-axis is time, the y-axis is frequency, and darkness represents intensity. Note that over much of the utterance the regions of greatest intensity are changing continuously.

Scientists use these mechanisms to study the details of events that occurred in the past. Phoneticians, for example, spend plenty of time contemplating spectrograms like figure 12.1. But what about animals (or people) in the field? They have to act immediately. What can they do to analyze and recognize events that only unfold over time? To think clearly about this, we need to first consider the kinds of information that will be of potential use to an animal.

Biological Time Animals (and humans) use temporal information for at least two general reasons. First, they use timing information to "recognize things," i.e., to relate events in the environment with previous experience of the objects and events. This includes objects like the presence of other animals, spoken words in human language, a banging window shade, the sound of a horse's gait, etc. Each of these "objects" imposes a characteristic structure on sound over time. Patterns that extend in time can often be usefully labeled by a name or by some "cognitive object." That is, temporal events can produce something rather like a "symbol" (van Gelder and Port, 1994). "It's a *waltz*." "She spoke *my name*." "It's *my mother* walking down the hall." A pattern extended in time, then, can cause a stable state to be entered into by the perceptual system, as a kind of recognition state.

The second main reason to use temporal information in sound is to support activity by the body. Ongoing perception directs action. The response of the perceptual system to a familiar auditory pattern will sometimes be to directly adjust the body to the input pattern itself. An animal can simply turn its head to face an important sound, or begin to intercept the object generating a sound, or occasionally to imitate another person's pronunciation of a phrase. Sometimes we even clap hands or dance to sound. So recognition of "things" is only part of the problem. An animal sometimes needs to bind its own real-time behavior to temporal events in stimulation. This frequently requires predicting the timing of future actions of another animal for some period into the future.

These are challenging functions. How can these jobs be fulfilled? And how many of these functions are directly supported by a spatial map of absolute time like figure 12.1?

Naive Time

What we call the "naive view of time" is simply the notion that biological time, i.e., the temporal information used by animals, is based on a representation of absolute time. In the realm of audition, it is manifest in the assumption that a critical early step in auditory processing of cognitive-level information must be to *measure time in absolute terms.* Typically, naive time models assume that humans or animals have access to real historical time. They apparently presume some clock that can measure durations directly in seconds. For auditory perception, the method usually implied is that the brain stores lists of spectrum-time pairs, i.e., a kind of buffer not much different from the sound spectrogram of figure 12.1. Such a display is often described as "short-term auditory memory."

This idea is widespread among psychologists, linguists, and phoneticians, and is probably assumed by most laypeople as well. It is supported by our intuition that we can remember recent sound events rather accurately. In addition, for many people, measurement of time in seconds and hours is the only natural way to think. If one assumes that every event has a location in time,—some angle of the clock at which it occurred—then locations and perhaps even durations in seconds may seem to be almost *intrinsic to the events themselves,* perhaps as intrinsic as the identity of the events that occur at each point in time.

Illustrative Models What is required for absolute time measurement is what we call an *assignment clock,* a device that moves (changes state) at a constant rate and supplies a unique label to describe each point in time. One also needs a set of descriptors, i.e., a kind of alphabet, for characterizing the set of possible events. Thus, for example, an audio tape recorder uses constantly rolling wheels to lay down values corresponding to the sound pressure over a very short time window on magnetic tape. In the sound spectrogram of figure 12.1, the descriptors are energy levels in a set of frequency bins over some time interval. Time (as long as it is greater than some Δt) is translated into a corresponding unique position.

Psychologists have explored the possibility of such a time-buffer for many years. In the 1960s "visual iconic memory" was discovered—a spatial image of the visual field in which objects are still raw and unidentified (Sperling, 1960). It was a level of visual memory from which a subject can select subportions for verbal description while unattended parts of the image are soon lost. This led to postulation by analogy of a short-term memory for sound (Crowder and Morton, 1969; Massaro, 1972; Baddeley, 1992), one that was sometimes called "echoic memory" (Neisser, 1967). The models actually proposed, however, did not resemble an echo at all. Since a real echo is an event in time, it seems that an echoic memory should be something that *replays* in time, like a tape loop that can be repeatedly scanned. But most theoretical models "cut" the tape loop so it can be examined all at once as an auditory "image."

One model for such a memory might be implemented rather like a postwar radar scope with a phosphorescent screen. These scopes displayed a decaying image of rain clouds or airplanes for a couple of seconds until the radial sweep rescanned that part of the circle and rewrote the image anew. Like the sound spectrogram, of course, this logically requires an assignment clock with its mechanism of constant motion to generate a spatial layout of the input sound spectrum for the past second or so, in order to serve as an auditory memory model. The spectrum just behind the sweeping radius would be the most recent. The sweep wipes out the decaying old information from the buffer. At least that is one way such a model might be implemented. However, without specifying any particular concrete mechanism, many models of auditory pattern recognition (including, e.g., Massaro, 1972, 1987; Klatt, 1980) have proposed a similar kind of short-term auditory memory lasting nearly a second that contains raw spectra straight from the auditory nerve. All meaningful auditory features are to be extracted from this representation. Durational cues, such as voice-onset time, are thus treated exactly the same as spectral cues—measured, apparently, by straightforward examination of the position of various spectral features arrayed along the time axis in short-term auditory memory. Recognition of temporal patterns is thus (naively, we would say) turned into a task that closely resembles recognition of visual patterns.

In the study of phonetics, time has posed recurring theoretical difficulties. Whereas linguistically motivated models of phonetics rely entirely on sequential order (Jakobson, Fant, and Halle, 1952; Stevens and Blumstein, 1978), phoneticians frequently found evidence that timing detail played an important role in speech perception and production (Lehiste, 1970; Klatt, 1976; Port and Dalby, 1982; Port, 1981). In one well-known controversy, Leigh Lisker and Arthur Abramson (1964, 1971) argued that voice-onset time, the time interval between the burst of an aspirated stop to the onset of voicing (as in the word "tin") was an example of a durational feature that was controlled by speakers and also employed by listeners in differentiating "tin" from "din." They claimed that the serial order alone would not properly differentiate these words; only a metrical measure would suffice. Thus, they argued, speech requires better measurement of time than simply the serial order of features (as proposed by Chomsky and Halle, 1968).

Of course, to serve as perceptual information, listeners themselves must somehow be able to utilize actual voice-onset times. Little was said by Lisker and Abramson (or anyone else) about a practical perceptual mechanism for this. But since phoneticians themselves just measure acoustic durations of speech by applying a ruler to sound spectrograms or from a computer screen, one must conclude that, to the extent that phoneticians consider such measurements relevant at all to the problem of human speech perception, they implicitly suggest that human subjects are also able to extract equivalent measurements. Thus far, however, the evidence for any ability to measure time in milliseconds is strictly circumstantial: it is clear that sometimes people are sensitive to quite small duration changes. But does this imply they measure time in absolute units?

Theories of speech production have also depended on naive time in some cases. Directly analogous to the perceptual models, hypothetical speech production processes are sometimes proposed that include a stage at which there is a list of letterlike phonetic segments paired with appropriate durational specifications. The buffer of these is then read out, from left to right, during actual speech production, and the gesture for each segment is executed so as to last just the specified amount of time. Thus Klatt (1976), followed by Port (1981), proposed specific "temporal implementation rules" that compute how long the phonetic states (i.e., various consonant and vowel segments) are supposed to last given their inherent segment durations and specific features of their context. In order for such numbers to serve as instructions, of course, a "motor executive system" must be assumed that is able to assure that the corresponding gestures do indeed last the correct amount of time. But there are many difficulties with this proposal (Fowler, Rubin, Remez, et al., 1981). To the extent that these are taken to be models of human behavior, they assume that durations in absolute units like milliseconds are intrinsically meaningful and interpretable. In short, temporal implementation rules are instances of naive-time models for motor control. Of course, many other approaches to motor control have avoided this pitfall and are based on dynamical models analogous to what we propose here for perception (Bernstein, 1967; Kelso, Saltzman, and Tuller, 1986; Saltzman and Munhall, 1989; Browman and Goldstein, 1989).

What we are calling the naive view of time, then, amounts to the assumption that time measured in milliseconds (a) is automatically available to a perceiving system and (b) serves as useful information in a system for motor control. Most researchers in both speech and other auditory patterns have focused attention on static problems—perhaps in part to avoid dealing with messy temporal patterns at all. Still there is a longstanding literature of research on specific temporal issues like rhythm production and perception (see, e.g., Fraisse, 1957; Michon and Jackson, 1985), but research on time has generally been treated as a backwater issue, not relevant to the major themes of psychological research. Perhaps in hope of attracting a little attention to the problem, one paper a few years ago bore the title "Time: Our Lost Dimension" (Jones, 1976). With a few notable exceptions (see, e.g., Povel and Essens, 1985; Jones, 1976; Watson and Foyle, 1985; Sorkin, 1987; Warren, 1993), including a large literature on speech perception, patterns that are distributed in time tend not to be viewed as important theoretical problems for perceptual theory.

Why is there a certain blindness to the unique problems of time in theories of psychology? One reason may be that it is simple to represent sound in a buffer based on absolute time measurements. Also, engineers have had at least some success in handling sound that way. For example, in speech recognition models, a buffer of audio input with discrete time labels (coded as spectral slices) was the basic data structure of the exciting Hearsay-II speech recognition system (Lesser, Fennel, Erman, et al., 1975). Although designed for engineering purposes, Hearsay-II has nevertheless served as a kind of archetypal speech perception theory for a generation of scientists. The model was based on standard structuralist ideas about the organization of a sentence: syntactic structure at the top, then a list of words, then phonemes, allophones, and acoustic cues. So, in order to recognize a sentence of speech, a second or two of audio signal is stored up in a buffer. Then a set of modules analyze various descriptions of the sentence, using phonetic, lexical, prosodic, and grammatical descriptors. These hypotheses are posted onto a "blackboard" with time as its *x*-axis, as shown in figure 12.2. Hearsay-II interprets the sentence all at once—only after the whole sentence has been presented.

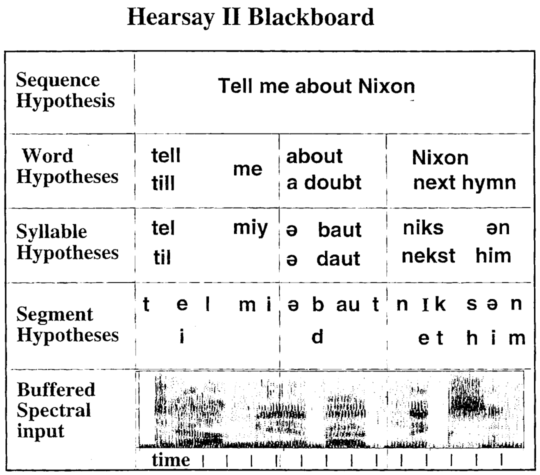

Figure 12.2 The Hearsay-II system stores the raw input in a buffer. Independent demons test hypotheses (such as that a particular phoneme is present in a particular temporal region), while simultaneously other demons look at the posted results of every other demon and create further hypotheses about other possible units (such as syllables). Thus gradually, a roughly simultaneous analysis is achieved for all levels of description for all sections of the utterance.

The behavior of the model over time (if we imagine running it continuously) would thus be to alternately collect data and then process it. One collects "enough" data (however much is needed for the problem at hand) and then crunches it. The structure of this model illustrates the basic form of many naive-time models of speech perception.

Connectionist approaches to speech using classic feedforward networks have had limited success at real speech recognition (Watrous, 1990; Elman and Zipser, 1988). This may reflect the fact that many connectionist models have continued the static tradition of dealing with time. For example, Elman and Zipser (1988) collect a syllable's worth of acoustic input into a buffer. Then the entire pattern is submitted to the neural network for analysis and recognition. Another current model for speech recognition, one that aims for high-quality performance, is the time-delayed neural network (TDNN) (Lang, Waibel, and Hinton, 1990). This model uses precisely controlled time delays to allow at least a syllable-length stretch of speech to be stored in a buffer that contains sampled absolute time as one axis. The recognition of syllable-length patterns takes place only after the whole syllable is present in the buffer. Of course, when researchers are solving practical problems, they should do whatever seems as though it might help. But cognitive perceptual models for speech and music perception, and so on have been copying features of these systems. Unfortunately, to do so is naive.

Problems with Naive Time There are two critical difficulties with the naive view of time surveyed above: (1) the lack of direct evidence for a temporal buffer and (2) the surprising lack of usefulness of millisecond measurements. If we are to rely on time labels (or physical positions) to record the past, we must depend on a highly accurate assignment clock (e.g., an audio tape recorder, sound spectrograph, digital-to-analog converter, video recorder, etc.). This clock assigns labels to descriptions of the energy layout of events. What evidence supports such a mechanism in humans or other higher animals? Note that our everyday measurement of absolute time is only made possible by various modern technologies that allow us to compare a real-world event with some device whose rate of change is presumed constant: a mechanical pendulum clock, the rotation of the earth, or the oscillation of a cesium atom.

The hypothesis of a spectrographic auditory memory makes at least three strong predictions. First, since the memory must have a limited duration, we should expect very good measurement of time for the duration of the memory, but then a sharp falloff in accuracy and perhaps greater reliance on relative durations for patterns longer than the short-term memory. Second, since it stores time only as absolute values, we should expect that patterns defined by relative durations should be more difficult to learn than ones defined in absolute terms. Similarly, subjects should need to be exposed to rate-varying productions in order to learn to recognize a pattern that is defined only relationally. Third, it should be fairly easy to recognize the absolute time alignment of unrelated events, e.g., "When, during the pronunciation of the word 'Indiana,' did the car door slam?" On any spectrogram of such a complex event, the relative position of the two events is obvious. As far as we can tell, for humans (and presumably other animals), *none* of these expectations holds. Absolute measurements are harder than relative ones for events at most time scales. Generalizing a pattern across a change of rate is easy and natural. Yet we perform very poorly at judging the lineup of unrelated events.[1]

The second major problem with naive-time models is this: time measured in seconds is simply the wrong kind of information for many problems. Listeners to environmental sounds, music, and speech have much less use for absolute time measurements than one might think. Both melodies and words (and many other kinds of auditory patterns as well) retain their identity even when their rate of production is varied. And changes in rate tend to change the durations of all the segments in a pattern uniformly. For example, knowing that some vowel in a word (or a note in a melody) is 150 ms in duration is, by itself, almost useless information regarding its identity and its role in the word (or melody). On the other hand, knowing its duration in relation to the duration of a number of (labeled) neighboring segments is very informative indeed (Port, Reilly, and Maki, 1987). Rather than an absolute time scale, what is much more useful is a scale intrinsic to the signal itself—a scale that will support local comparisons, such as durational ratios (Port and Dalby, 1982). Internal perceptual mechanisms may be able to lock onto some period and measure relative durations as phase angles and even predict future events in the pattern.

Other Ways to Measure Time

We propose that ecological patterns have at least two kinds of time information that are "weaker" than absolute time, but nevertheless very useful: serial order and relative duration.[2] Serial order is a topic with a long and well-developed history. The analysis of strings of symbols is the basis of much of computer science as well as linguistic theory. Relative duration, however, has received much less attention and few mechanisms have been explored for its extraction and description.

We hypothesize that human cognition, like the cognition of other less sophisticated animals, does not have a completely general-purpose store of raw acoustic information that is created in advance of pattern recognition. Instead, recognition, i.e., a labeling process that depends on extensive previous experience, precedes the generation of whatever buffers there may be of events in time. The kind of short-term auditory memory we propose contains labels or "names" for analyzed microevents. Each event contains its own time specification. No abstracted time scale may exist at all. The unfortunate consequence of this state of affairs, is that familiar patterns each have an appropriate temporal representation, but if listeners are presented with a completely novel pattern (not containing obvious periodicities) or if several familiar patterns overlap in time, listeners have only very weak resources for representation of such complexes (Port, 1990).

Since absolute time representation must be ruled out as a general method, despite clear evidence that animals and people are very sensitive to many temporal structures, what other possibilities are there?

Serial Order The weakest descriptive scheme for time is just to specify the serial order of events. Such a description is what linguistic models provide: the standard European alphabet, widely used for orthographic writing systems, is a classic tool for linguistic analysis in the slightly modified form of the phonetic alphabet. Words are made up of phonemes, serially ordered like beads on a string but with no durational properties (i.e., the only measure of length is the number of segments). Sentences, in turn, are viewed as nothing but serially ordered words. Our commonsense understanding of how events are ordered due to a relation of cause and effect also leads to expectations of serial order: a sudden squeal of brakes causes us to expect the sound of a collision, thunder follows lightning, and click follows clack in the many whirligigs of modern life. Serial order may be noncausal as well: when one hears a shoe drop on the floor upstairs, one may expect to hear the other one after some unpredictable delay.

Early speech recognition models which grounded measurement in absolute time ran up against a myriad of problems due to the intrinsic variability of speech timing. The most successful of these systems modeled speech as a series of ordered states using techniques like "dynamic time-warping" to get rid of much of the absolute information. Still, mere order, with time measurement achieved by counting segments, will not do the job for many important environmental events. If there is a periodic structure of some sort in the input signal, then an effective auditory system can exploit that regularity both to predict and to describe.

Relative Duration Relative duration is just the comparison of one duration with another. Like other ratios, it is dimensionless. We may arbitrarily select one unit as a reference unit. If the reference time unit is extremely regular, like the motion of our planet relative to the sun, then relative time approaches equivalence to absolute time. But other, context-sensitive, reference units are also possible—a period detectable from the signal. We can enumerate periods just as well as seconds. Instead of fractions of a second, phase angle can be measured with respect to the reference period. Then if the rate of the input pattern changes slowly, our scale can remain calibrated. The difference between this intrinsic referent and absolute clock time is enormous because for many ecological events, a relative scale of time is much more useful.

The fundamental reason for the value of relative duration measurements is simply that many dynamic events in the environment that are functionally equivalent (i.e., have the same meaning) can occur at a range of rates: characteristic animal or human gaits, musical rhythms, songs, engine noises, the swaying of tree limbs, and, of course, spoken words. If you want to recognize a waltz rhythm, it should not matter much what the rate of the rhythm is in milliseconds per cycle. This property is acknowledged in the standard notation system of Western music which employs a notational variant of phase angle for time measurement: thus, in a 4/4 time signature, a half-note represents the duration of π radians (relative to the "measure"). Indeed, most forms of music around the world are constructed around such periodic, partly rate-invariant hierarchical structures.

But a complex signal may contain subparts, whose duration relative to the signal rate is of importance. For example, it is clear that relative timing, not just serial order and not absolute time, plays a major role in the information for speech timing (Port, Dalby, and O'Dell, 1987; Port, 1981; Port and Dalby, 1982; Lehiste, 1970). A well-known example is the syllable-final voicing distinction in English and other Germanic languages. One of the major cues for the distinction between pairs such as *rabid* and *rapid* or *camber* and *camper* is the relative duration of the vowel to the postvocalic stop consonant or consonants. This is more naturally expressed with reference to the syllable period, rather than the second (Port et al., 1987; Port and Cummins, 1992). A satisfactory account of speech perception requires time measurement that is *more powerful* than just serial order, but clearly must be *less powerful* than absolute time in seconds.

Need for New Approaches

Thus far we have argued that the widespread view of time as somehow naturally assigned in seconds is not usually an appropriate approach to the study of perception of temporal patterns by animals. It presumes neurological mechanisms for which little direct evidence exists, and does not provide the most useful description of information without further arithmetic processing that would throw away the absolute information obtained with such difficulty. As an alternative, one can analyze time with just serial order. This has been attempted many times and seems to be adequate for some aspects of problems like grammatical syntax. However, such an approach leaves many phenomena unaccounted for. For example, what about events that are regularly periodic? Serial order contributes nothing to understanding temporal measurement of this type. It is not sufficient merely to get the notes of *The Merry Widow* in the right order if the note durations vary randomly. And if the note durations are specified symbolically (as in musical notation), how can these be accurately implemented for production or accurately recognized in perception? How do listeners obtain or use this temporal information? What kind of mechanisms can listeners employ to be able to measure all the major classes of temporal information?

Our hypothesis is that listeners employ a bag of temporal tricks. As they gain experience with their auditory environment, they develop a variety of mechanisms for capturing spectrotemporal patterns—the specific ones that occur frequently. To a significant degree these structures are self-organized (Anderson, 1994) and do not require explicit tutoring. Wherever possible, these mechanisms will exploit any periodicity in stimulus patterns. If none can be detected, then serial order may have to suffice—but in any case, temporal structure is learned as part of the patterns themselves, not as an independent abstract dimension. The best way to study these mechanisms in our view, is to simulate them computationally using simple dynamical models and then to compare qualitative properties of performance with human or animal data. The dynamical systems we propose are orders of magnitude simpler than the dynamics of real nervous systems and, consequently, could be plausibly implemented by biological systems. In the following sections, we suggest several general methods for extracting useful temporal information, both with respect to serial order and relative duration.

12.3 MEASUREMENT MECHANISMS

The two methods of serial order and relative duration are closely related to S. S. Stevens's notion of an ordinal scale vs. an interval scale (Stevens, 1951) and, like them, they are conceptual types of measurement. Any actual achievement of such measures in a nervous system may involve a wide range of mechanisms. In this section we review some methods that will allow recognition of both serial order and relative duration. As we shall see, both methods depend on the behavior of internal dynamic models to keep track of time in the perceptual system.

Recognition of Serial Order

Some models that identify the serial order of elements in a temporal sequence were developed by those working on speech recognition in order to overcome the problem of invariance of patterns across changes in the duration of individual components as, for example, due to a change in rate or emphasis. To achieve this, it is useful to factor out as much durational information as possible, focusing on transitions from one element to the next. The first model, the finite-state machine (FSM) has certain strengths. The same ideas appear again in slightly disguised form in the hidden Markov model. We will show how a dynamical system can emulate an FSM that runs in real time and what advantages it possesses in dealing with continuous signals.

Finite-State Machines The traditional mathematical system that recognizes and classifies sequences of events is called a finite-state machine (see Pollack, chapter 10, for further discussion of these systems). An FSM consists of a set S_i of states, including two privileged kinds of state, S_1, the start state, and S_{accept_i}, a subset of S, containing one or more "accept" states. If the FSM is in an accept state, it has successfully recognized the sequential pattern for which it is specific. Transitions between these states are defined in a transition table, in which entry e_{ij} is the state to which the machine is set when it is already in state S_i and receives the input I_j. Transition is therefore dependent *only* on the state at the previous time step and the current input. The only memory the FSM has for earlier inputs therefore resides in the state of the machine. Figure 12.3 shows a simple FSM which has four states and an input vocabulary of three symbols, *A, B, C*. The transitions that are labeled are the only ones that may appear if the machine is to recognize a sequence. All other possible combinations of input and state lead to an implicit "reject" state, from which there is no return. The illustrated FSM will accept, among others, the sequences *AABBBC* and *ABCCCC*, but will reject *AAABBA* and *AAACCCBBC*.

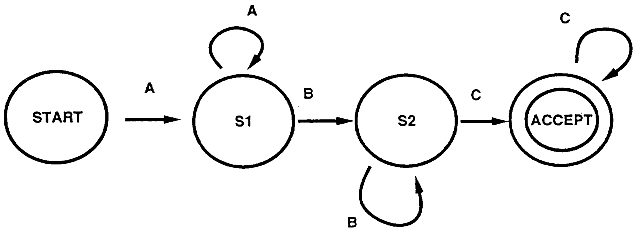

Figure 12.3 A finite-state machine which recognizes the sequence *ABC*. It makes transitions from state to state only when a particular letter is received. Thus only a *B* will allow the transition from *S1* to *S2*. Transitions other than those shown cause the machine to reject the sequence. The final state (with a double circle) is the only "accept" state.

One of the most successful of the first generation of automatic speech recognizers was Harpy, a system based on a large FSM with some 15,000 states. The Harpy system was the most successful entry in the 1971–1976 speech recognition project sponsored by the Advanced Research Projects Agency of the Department of Defense (Lesser et al., 1975; Klatt, 1977). It contains a hierarchy of nested FSMs plus a search procedure to identify the path through the space of possible input sequences with the least total error. Figure 12.4 shows schematically how these FSMs, here represented as networks or graphs, are layered. Phoneme recognition networks scan the raw acoustic buffer, trying to identify individual phonemes. These in turn serve as input to a lexical-level FSM. One of the principal advantages of a Harpy-like system was the fact that no time normalization was required. Just as in our example FSM above (see figure 12.3), each state has self-recurrent transitions. Thus, if a single element (phoneme, word, etc.) is repeatedly presented (or presented more slowly), the network does not move from its current state. Thus in principle, one could stretch one segment (say, the *a* in *about*) for an indefinite number of time steps and Harpy would still recognize *about*.

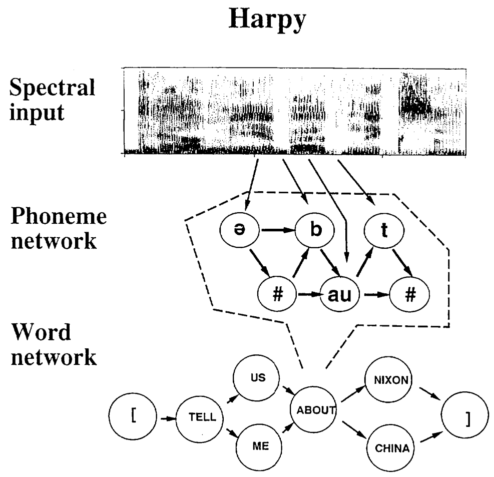

Figure 12.4 A schematic representation of the Harpy system. Hierarchies of finite-state machines scan the acoustic input and recognize units at ever greater scales, from phonemes to words to sentences. Here, a set of phoneme recognizers are fed spectral input. As each in turn recognizes "its" part of the input, it outputs its result to a lexical unit which pieces together words. The order of the spectral slices yields phonemes, the order of the phonemes yields words, and the order of the words yields grammatical (acceptable) sentences.

This contrasts sharply with Hearsay-II, discussed above, since Harpy employs no direct representation of time—merely the order of events. Of course, if such a model is refined to allow recognition of additional variant pronunciations, the size of the transition table for the FSM will increase exponentially. If the signal being studied is not easily reducible to a reasonably small number of states or features, then an FSM model rapidly grows to unmanageable size. This problem of the proliferation of states applies not only to FSMs like Harpy but to any of the more sophisticated Markov models, which are ultimately based on the FSM. We will look briefly at hidden Markov models, which represent the current state of the art in speech recognition, and then show how a simple dynamical model can circumvent this problem of exponential growth in complexity.

Markov Models Finite-state machines are a straightforward way of recognizing sequences and can be used for producing sequences as well. In the example FSM given above, the sequence *ABC* could be presented at any rate (i.e., any number of *A's* followed by any other number of *B's*, etc.) and *still be* recognized by the FSM since it would be guaranteed to reach a goal state. Time has been factored out completely. In dealing with real-world signals, however, where the exact sequence of possible states is not known for certain, the model needs more machinery. In many successful models, the FSM is augmented as follows: it is assumed that the unknown signal has been generated by an FSM which outputs a symbol in each state. An attempt can now be made to reconstruct the generating model by a process of inference. In order to do this, probability distributions are obtained for both the outputs associated with each state and with the transitions from one state to the other. The model that is inferred is known as a hidden Markov model. The assumptions of the FSM have been retained, since transitions depend only on the previous state and the current input, but has been augmented by probabilistic transitions between states. It is hidden because the deterministic FSM has been replaced by a best guess, in which the outputs are generated stochastically and cannot be known with certainty. Hidden Markov models are at the base of many contemporary speech and temporal pattern recognition models (e.g., Lee, Rabiner, and Pieraccini, 1992; Lee, 1992). As mentioned above, they too may run into the problem of state proliferation if the underlying signal is not easily reducible to a small number of relatively steady states.

Simple Dynamic Memories In the last few years, several new approaches to speech recognition have emerged within the area of artificial neural networks. Most innovative has been the use of recurrent networks that process a small amount of external input at a time and retain information about the past only in the particular internal activation state of the fully connected units. In the best cases, these networks have outperformed the hidden Markov models, with the advantage of requiring no domain-specific knowledge to be encoded by the programmer (Robinson and Fallside, 1991). They have been applied to a number of problems in speech recognition and appear to hold promise for many kinds of pattern recognition. We present a recurrent network model of our own that is similar in many ways to an FSM recognizer, at least over a limited range of stretching or compression of time. The model nevertheless has some significant advantages that come from having a continuous state space rather than a discrete one.

One of the many tasks which recurrent networks have proved to be good at is the emulation of FMSs (Pollack, 1991; Das, Giles, and Zheng Sun, 1992; Cummins, 1993). They are fed as input the same finite string of symbols as an FSM and the output is trained to reflect the distinction between accepted and rejected sequences. Rather like FSMs, the properly trained network will recognize the same sequence of elements, despite considerable variation in the rate at which they are presented, e.g., *AAAABBCCC* will be recognized as being the same sequence as *ABC*. This "normalization" is perhaps surprising, since, during training, the network may have seen each sequence presented only at a single rate. The generalization across rate changes was obtained "for free." In the remainder of this section, we look more closely at the dynamics of the trained network and see how this "rate normalization" is achieved.

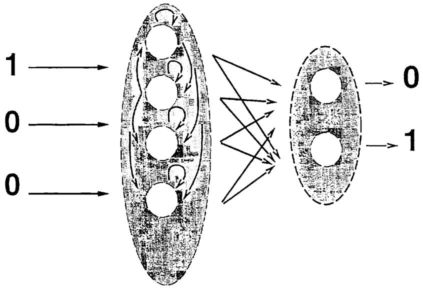

Figure 12.5 Schematic diagram of a recurrent network that accepts (sensory) input on three input lines, processes the input, and outputs a binary tuple. Each unit sums over the input it receives from the input lines and from other units, computes an output value, and passes it on. Because of the recurrent connections, information about previous inputs is retained in the current state of the recurrent units.

A recurrent neural network, like the one shown in figure 12.5, is a processing model comprising several external input lines (from "sensors") feeding to a layer of fully connected processing units. In this system, which we call a simple dynamic memory, each unit collects input from all the external lines and from the other units. At each time step, each unit sums over all its inputs, performs a simple squashing transformation on the sum, and outputs the result. Thus the activation equation is

$$y_i(t + 1) = squash[\alpha y_i(t) + \sum w_{ij}y_j + input + bias] \quad (1)$$

for connections from unit j to i, α is the decay rate, the bias serves as a threshold, and *input* refers to external (sensory) input only. Some units of the network are designated output units and the outputs on these represent the response of the model to the input (see figure 12.5). The input here is the external stimulus fed to some nodes. Because of the recurrent connections, information from previous inputs remains implicit in the state of the recurrent units. Gradient descent training procedures using a teacher signal for just the output nodes is employed (Williams and Zipser, 1989; Anderson and Port, 1990).

We have noted that the memory of a recurrent network arises from the change in internal state after each vector presentation. The dynamics of the trained network can be studied by looking at the trajectories followed by the network in the state space of unit activations. Assume that a network is trained to recognize the sequence *ABC* and distinguish it from, among others, *BAC* and *CAB*. The network signals recognition of *ABC* with an output node which is off (activation = 0) until the last element of *ABC* is seen, at which time it switches on briefly (activation = 1). The state space can be divided into two principal regions, the hyperplanes defined by *outputnode* = 0 and

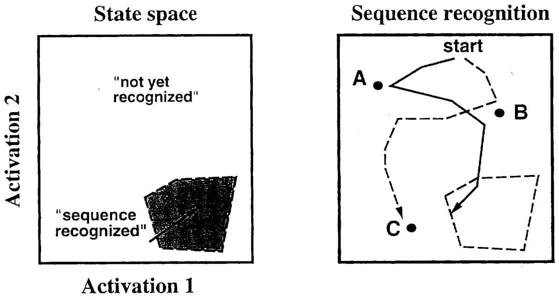

Figure 12.6 (*Left*) The state space, here, for simplicity, illustrated as being two-dimensional, is partitioned into a large *not yet recognized* area and a *sequence recognized* region. When the system trajectory enters the recognition region, it signals sequence identification. (*Right*) For each possible input, *A*, *B*, *C*, there is an associated point attractor. As input changes (e.g., from *A* to *B*), the system trajectory is redirected toward the new attractor. Only the trained sequence brings the trajectory through the recognition region (since it was the learning algorithm that located the recognition region in just the right place). The solid line is the trajectory on presentation of *ABC*, and the dashed line is *BAC*.

outputnode = 1, associated with nonrecognition and recognition, respectively. This is illustrated in figure 12.6 (left). If we present the network with a continuous, unchanging input, it rapidly settles into a steady state. Thus a global point attractor can be identified for each of the possible input vectors, including the zero vector (no input). Figure 12.6 (right) illustrates how this system is able to distinguish the sequence *ABC* from all other sequences of *A's*, *B's* and *C's*, such as *BAC*, etc.[3] Assuming that the system is reset to some neutral state between sequences (marked *start*), the trajectory corresponding to the system evolution can be visualized as always starting in the same area of state space. As long as the first element, *A*, is presented, the system gradually approaches a point attractor specific to that input. Once the input changes to *B*, the attractor layout also changes and the system trajectory changes course toward the new attractor. The task of learning to identify a sequence now amounts to insuring that the trajectory passes through the recognition region if and only if the sequence to be identified has been presented. This learning can be based on either tutoring or self-organization, but it must be based on actual experience with the temporal patterns.

We can now illustrate how this general model, the "simple dynamic memory," handles varying rates of presentation. Figure 12.6 (right) shows the system trajectory as jumps in discrete time, since each sequence element is presented for an integral number of clock ticks, *AAABBBCCC*.... The trained network can now be presented with the same sequence, but at a different rate, and it will still successfully distinguish among targets and distractors. This is illustrated in figure 12.7 (right). The two trajectories illustrated are for presen-

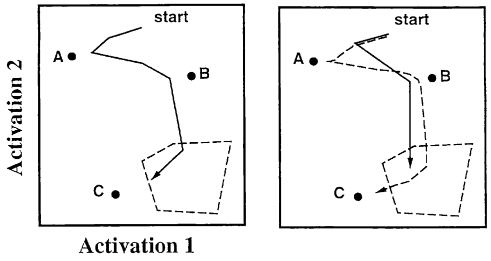

Figure 12.7 (*Left*) System trajectory taken by the network trained on *AABBCC* when the target sequence is presented. (*Right*) Trajectory of the same network when the sequence is presented at novel rates, both faster (*ABC*, solid line) and slower (*AAAAABBBBBCCCCC*, dashed line). The trajectory still passes through the recognition region.

tations of *ABC* (*solid line*) and *AAAAABBBBBCCCCC* (*dashed line*), which, assuming constant sampling of a continuous signal, represent faster and slower presentations of the sequence $A^nB^nC^n$ (for small n). Despite the variation, the underlying dynamics are nearly the same, and the trajectory remains qualitatively unaltered. In each case, the trajectory still passes through the recognition region. Thus, like the FSM, this system intrinsically ignores variations in rate.

There are additional parallels between this model and an FSM. The part of state space we call the recognition region corresponds to an "accept" state, while the individual attractors, together with their vector fields, correspond to individual states. Unlike the FSM approach, the dynamic memory will ultimately break down if patterns are presented too slowly. As the state of the system gets closer to the fixed point of the current input (after the input is repeated many times), it becomes increasingly difficult to differentiate the effects of previous events since there will always be limited precision in activation space. However, this type of solution has its advantages too. In particular, it generalizes to continuously varying input, without growing in complexity. It is therefore more suitable for signals which vary smoothly and are not easily reducible to a small number of discrete states.

In order to show this property imagine an input set of at least two orthogonal vectors, each of which has a distinct point attractor. As input varies continuously from *A* to *B*, the attractor itself may move smoothly from the point associated with *A* to that associated with *B*, as shown in figure 12.8. The continuous nature of the state space allows smooth interpolation between attractor regimes (cf. chapter 5, by Beer, which illustrates a model with this feature). This behavior is without parallel in the perfectly discrete FSM. The induction of such a dynamical system presents a technical problem, as there is no guarantee that the dynamics will always remain as well-behaved as observed thus far (see Kolen, 1994). For example, attractors may split in two, producing bifurcation. Or they may be more complicated than simple point attractors and undergo catastrophes. However, recent work in the induction of general FSM-like dynamic systems has been encouraging (see Pollack, chapter 10; Das et al. 1992). With well-behaved dynamics, simple dynamic memory offers the possibility of an FSM-like approach to the recognition of serial order which generalizes to continuous dynamic signals such as speech.

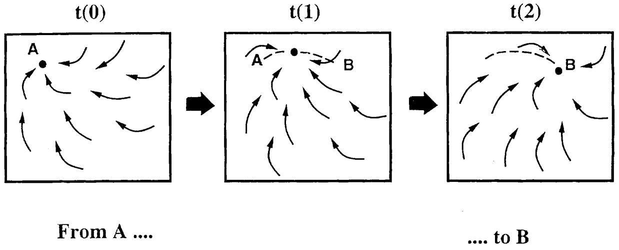

Figure 12.8 Continuous interpolation between attractor regimes is possible with a dynamic system which emulates a finite-state machine (FSM). Unlike the FSM, the system does not increase in size or complexity as we generalize from a small set of discrete inputs to an unbounded set of continuous inputs.

We have shown that serial order is a kind of information about events in time that a simple network can recognize and identify. Although more sophisticated models of serial pattern recognition exist (Grossberg, 1986, 1988; Anderson, 1994), the simple dynamic memory illustrates one method by which problems resulting from variation in rate of presentation can be overcome. The biological correctness of the method described here, it should be noted, has not yet been demonstrated.

Recognition of Relative Time Patterns

The simple dynamic memory model of unstructured fully connected nodes is capable of recognizing a familiar sequence of ordered elements, independent of the rate of presentation. In doing so, it loses (or ignores) absolute time. Many problems, however, depend critically on more powerful measurement of duration than serial order. Simple dynamic memories will not suffice for these patterns. But durations of importance to perception of the environment are frequently relative; i.e., they could be adequately measured with respect to salient intervals within the signal itself. Relative durations are critical for animal gaits, periodic bird and insect sounds, and so forth. In Western music, the basic time unit is the beat. In speech it will be the syllable period for some measurements and perhaps other prosodic units (like the foot or mora) in other languages (Port et al., 1987; Anderson and Port, 1994; Lehiste, 1970).

Having such a periodic standard, be it only a slippery and flexible one, can provide a practical means for measuring duration, evaluating rates of change and for entraining activity to an external signal (as in playing music or dancing to it), and for defining rhythmic structure. In order to detect rate and thus calibrate duration, a mechanism which can rapidly entrain to the period of a signal is useful. In fact, it is not too difficult to design such a device as long as a distinctive entraining event can be identified, i.e., an event that defines phase zero. A practical "adaptive oscillator" was developed in our laboratory following earlier work by Carme Torras (Torras, 1985; McAuley, 1994a,b; Large and Kolen, 1994). Once more, this can be achieved with a simple dynamical model. We describe the model in some detail here, because of its importance to our central theme.

First we should distinguish the adaptive oscillator from the more familiar case of a pair of continuously coupled oscillators, in which the phase of each oscillator continuously influences the periodic behavior of the other. The net result is an oscillation frequency for the coupled system that is determined by the coupling strength and the intrinsic oscillation frequencies of each oscillator. As soon as the oscillators become uncoupled, each independent oscillator immediately resumes its intrinsic oscillation frequency.

In contrast, the adaptive oscillator mechanism augments a pair of pulse-coupled oscillators in which only the phase information at each input pulse of one of them is available to influence the oscillation frequency of the other. Based on this phase information, the adaptive mechanism makes a more durable internal change to the intrinsic frequency of the oscillator. Thus, specifically, an input pulse occurring somewhat before the oscillator would spontaneously fire causes two things to occur. First, the oscillator spikes immediately and begins a new cycle. That is, phase is reset to zero. Second, the oscillator makes a small change in its own natural period, so that it more closely resembles the period of the input. In this way, the oscillator will quickly entrain itself so as to fire at the same period as the input signal (or at some integral multiple or fraction). If the input signal ceases, a slowly acting decay process causes it to drift back to its original period. This process of synchronization and decay is based on a gradient descent procedure described in more detail below. It is easy for the model to adapt to tempos that are near its preferred rate, but increasingly difficult to adapt to tempos that are significantly faster or slower. These may result in entrainment to harmonic ratios other than simply 1 : 1 or 1 : *n*. Ratios such as 2 : 1 or more exotic ratios like 5 : 2 can also be attractor states of the system.

The Adaptive Oscillator Model The specific adaptive model that we will look at is the adaptive simple harmonic oscillator as described by McAuley (1994a). This model has been applied to psychophysical data on the ability of listeners to discriminate small changes in the rate of isochronous auditory pulses (McAuley, 1994b).

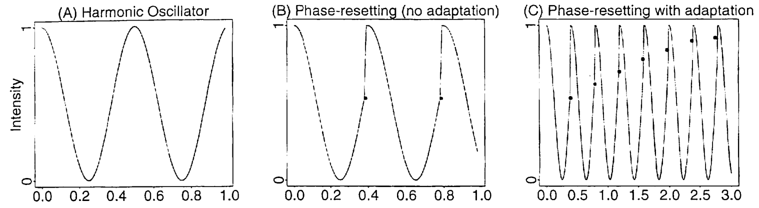

Figure 12.9 (*A*) Two periods of a 2.0-Hz harmonic oscillator. (*B*) Input pulses are now added to the harmonic oscillator every 400 ms, a shorter period than the intrinsic rate. Each input pulse now causes the oscillator to spike and phase-reset to 0. Output values, equal to the activation of the oscillator when it spikes, are marked by the dot at each phase reset. (*C*) Fast-acting synchronization is applied to the oscillator. Note that output values at each phase-reset continue to increase, providing a measure of the degree of entrainment. The output approaches the value of 1 as an attractor.

The preferred period of the oscillator is based on a periodic activation function, which in this case is simply a cosine function, scaled to oscillate between 0 and 1 (figure 12.9A):

$$\phi(t) = \left(1 + \cos\left(\frac{2\pi t}{\Omega(n)}\right)\right)\Big/ 2 \tag{2}$$

The oscillator period $\Omega(n)$ is initialized to some preferred period: $\Omega(0) = p$. Each time $\phi(t)$ reaches or exceeds a threshold, θ, set to 1, the oscillator generates an output spike. In the absence of external input, this happens at the end of each period (at phase = 0). We now add periodic input (figure 12.9B) to $\phi(t)$, producing a total activity that is the sum of the basic activation function plus the input:

$$a(t) = \phi(t) + i(t) \tag{3}$$

Again, each time the threshold is reached, the oscillator fires, but now it is firing before the end of its period. On firing, we introduce a discontinuity by resetting the phase to 0. Figure 12.9B illustrates the effect of adding the input to the intrinsic activation (the threshold θ is 1.0). Each time input occurs, there is a jump in activation and phase is reset to 0. It is useful to define an output spike function $o(n)$, by letting $o(n) = \phi(t)$ at the point in time when i arrives. This is marked with a dot in figure 12.9B and C.

How does adaptation or entrainment work? If an input happens to arrive exactly in phase with the intrinsic period, it will have no effect, since the oscillator fires then anyway (e.g., at $t = 0.5$ second in figure 12.9A). If it arrives at any other time, however, it will force the oscillator to spike earlier (as shown in figure 12.9B). This phase difference provides information which is used to define a spike-driven gradient-descent procedure which synchronizes or entrains the spontaneous spiking behavior of the oscillator with rhythmic aspects of the input pattern. The result of this adaptation can be seen in figure 12.9C, where the amount of jump when inputs occur can be seen to decrease as the period of the oscillator comes into line with faster frequency of the input. The synchronization error is defined by squaring the distance in time between input-forced spikes and spontaneous spikes. This is simply the squared difference between the threshold θ and the spontaneous activation $\phi(t)$, scaled by the input:

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

To minimize the discrepancy between the oscillator's initial spikes and the forced spikes, the oscillator's period $\Omega(n)$ is adapted by a small fraction α that is negatively proportional to the partial derivative of the synchronization error $E(n)$ with respect to $\Omega(n)$:

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

In this way, the oscillator adjusts quickly to the faster or slower frequency of a signal that excites it. One can also arrange for a decay process that will cause the oscillator, once adapted away from its initial frequency, to gradually return to that frequency, as a "preferred frequency." This can be done by including a term in the update rule that pushes the adapted period back toward its preferred period p. This decay should be quite a bit slower than the process of adapting to a periodic input.

Simulations have shown that these oscillators can entrain rapidly (largely within two or three periods of the periodic input). They are very robust to noise and occasional missing inputs since the decay rate is slow compared to the entrainment process. If the input is not perfectly periodic, but varies somewhat in frequency from one period to the next, the oscillator will still closely track the variations and should allow a good guess about the time of the next input.

Measuring Time as Phase Angle Some useful features of this oscillator model include its behavior in noise and when occasional inputs are absent. If there are irregular or noisy inputs, the oscillator will tend not to be affected. Because the entrainment process takes several oscillator cycles, irregularly spaced inputs tend to cancel out one another's effect, while periodic input will quickly cause the oscillator to approach the input frequency.

If an oscillator has a preferred period that is very different from the input signal, it may also entrain at ratios of, e.g., 2:1 or 3:1. These other ratios allow a single complex pattern to entrain a number of oscillators for different periodic components of the pattern: some at the "measure level" and others at the "beat level," etc. By setting up a number of oscillators with a range of preferred periods, it is possible to use the different entrainment ratios to extract the hierarchical rhythmic structure of the input. Some units will entrain at the most basic level such as the musical beat (or the fundamental frequency of voiced speech), while slower oscillators will pick out larger metrical units such as musical bars (or, hopefully, prosodic units like the syllable or foot). Such a hierarchical structure has been demonstrated in simulations in which a bank of oscillators with a range of preferred periods were exposed to two kinds of rhythmic inputs (McAuley, 1994a). The first was a waltzlike rhythm (*Dah dit-dit, Dah dit-dit*) and the second a two-beat measure (*Dah dit, Dah dit*). In each case, some oscillators in the simulation entrained at the measure level, locking into the strong beats only, while others picked out the individual beats. Those entrained to the beats are therefore measuring events that occur at fixed phase angles with respect to the larger, measure-sized cycle—either thirds or halves of the measure-size unit. Note that modest changes in the rate of the entire hierarchical pattern will not disrupt these relationships. This is a simple example of time measured purely as phase angle, rather like musical notation. The absolute duration of the period is irrelevant to the phase angle time measurements as long as regular periodicity is maintained.

This mechanism entrains rapidly to underlying periodic structures in a signal despite noise, missing beats, and slow variation in rate. Further development of adaptive oscillation should allow the measurement of time as phase angle under a broad range of conditions. All that is required is an unambiguous "start pulse" that must be supplied by a preprocessing system. Measuring relative durations without prior measurement of absolute time is now a possibility. In the case of speech, for instance, we hope it will prove possible eventually to entrain directly to the roughly periodic syllable and foot-sized units in speech. Phonetically relevant vowel durations might then be expressed in relation to these entrained periods rather than in terms of absolute values like milliseconds. By quickly and adaptively identifying and tracking the periodicities intrinsic to the speech signal, useful measurements of duration that are robust under changes in rate of presentation may be possible. Sounds that are merely quasi-periodic abound in nature and are by no means restricted to speech. Anywhere oscillation occurs—after striking, rubbing, blowing, dripping, as well as in animal gaits etc.—the signal may display an underlying period which might be exploited to scale the measurement of its subcomponents or other associated events. Adaptive oscillation offers a plausible mechanism for description of such temporal patterns, both as a model for neural mechanisms in cognition and potentially for engineering purposes as well.

12.4 CONCLUDING DISCUSSION

In this chapter we have highlighted the problem of recognizing auditory patterns in time. We claim that the naive view of time, despite its widespread employment, is not a useful model of any process in human audition. Events do not come with time stamps on them, nor does human audition supply them. It is difficult to imagine how there could be any direct representation of time either using labels marked in seconds or by translating time into physical distance (for anything beyond extremely short delays). Apparently nothing resembles a sound spectrogram in the auditory system. Thus the exploitation of durations in the acoustic signal is made possible by the use of the serial order of known patterns and by measuring duration relative to some predictable interval—not by measuring absolute time. We have made some first steps toward development of two models of auditory processing in audition that may simulate human performance for periodic patterns and sequences at the cognitive time scale. Both of the general methods described here are formulated as dynamical systems. In both cases, the behavior of these systems over time is exploited to keep track of location within a temporally distributed pattern. And in both cases, predictable features of the stimulus itself provide the yardstick for the measurement of time.

These methods are simple enough that we can imagine them implemented in many ways in the auditory system, but, of course, each method has certain apparent drawbacks. The method of simple dynamic memory for sequence recognition, for example, may offer rate invariance for free, but it requires actually learning an inventory of individual patterns. The system can only represent the serial-order structure of events it is familiar with. In our view, this property in no way disqualifies it as a model of processing in animal auditory systems. After all, most of the sounds we hear are, in fact, very similar to sounds we have heard before. Most animals live in environments in which the same kind of events recur. There may need to be a large inventory of auditory events, but any inventory is still minute compared to the space of possible frequency-by-time auditory patterns. Indeed, it is known that if listeners are presented with very novel yet complex auditory patterns, their ability to compare them or make judgments about their internal structure is astonishingly poor (see, for example, Watson and Foyle, 1985; Espinoza-Varas and Watson, 1986). Only practice with a specific set of novel patterns makes detailed comparison possible if patterns are both complex and completely novel.

In fact, given typical auditory ecosystems, this "drawback" of requiring familiarity with the patterns would have the practical advantage that when several familiar events happen to overlap in time (e.g., a spoken word and the slam of a car door), an auditory system that is able to represent only the learned set of patterns should automatically do "auditory scene analysis" (Bregman, 1990) and parse the complex into its familiar components.[4] Since the serial order of the subcomponents of familiar patterns were learned independently for each pattern, the temporal alignment between the two distinct events will, however, not be well represented. It would be very difficult to say which phonetic segments in the word coincided with the door slam. This accords with our intuition as well as with experimental results (see Port, 1990 for further discussion).

The measurement of relative duration, that is, measurement of a durational ratio between some event and a longer event, is useful for many kinds of sound patterns. Description of duration as an angular sweep of phase within a pattern of known period depends on predicting the duration of the longer event. The adaptive oscillators described here offer a way to do this when the input signal contains salient periodic events that can trigger phase resetting and period adaptation. Adaptive oscillators are quite simple to arrange neurologically, but obviously, to apply this mechanism to very complex auditory structures like speech or music will require (among other things) some highly sophisticated preprocessing in order to supply triggering signals for events of just the right sort. Adaptive oscillation should be a useful mechanism for analysis of many different kinds of environmental events and may be embedded in many places within a general auditory system.

One implication of the employment of adaptive oscillation for handling unfamiliar patterns should be that a pattern of clicks, say, that lack regular periodicity (e.g., with random spacing in time) will be much more difficult to remember or to differentiate one from another, than patterns of periodically spaced clicks. This has been shown to be the case (Sorkin, 1987; Povel and Essens, 1985). More subtly, if subjects listen to a series of several clicks and then try to determine if a second series of clicks has the same rate, performance improves as the number of clicks in each series increases from, say, 2 to 8 (after which there is no further improvement). This follows naturally from the hypothesis that more clicks permit closer adaptation of the perceptual oscillator to the input pattern and thus better discrimination (McAuley, 1994b).

In conclusion, then, it can be seen that the kind of auditory pattern recognition system we envision must be customized for a particular auditory environment. It is a system that organizes itself to construct a large inventory of special-purpose recognition mechanisms appropriate to the inventory of acoustic events that have relevance to the organism. These recognition mechanisms can not simply be part of long-term memory, or part of a system that analyzes the contents of a general-purpose, spatially arrayed short-term memory (like a sound spectrogram). Instead, we propose that these mechanisms themselves provide the first level of auditory memory. On this view, low-level auditory recognition and low-level auditory memory are collapsed into a single system that responds in real time to sound as it occurs. This system does not rely on a universal clock or other representational mechanism for absolute time. It consists, basically, of a bag of dynamical tricks that enable an animal to deal with the dynamically generated sound patterns.

ACKNOWLEDGMENTS

The authors are grateful to Sven Anderson, Kenneth deJong, Michael Gasser, Gary S. Kidd, Catherine Rogers, Tim van Gelder, and Charles Watson for helpful discussion of these matters and for comments on earlier drafts of the manuscript. This work was supported in part by the Office of Naval Research, grant number N00001491-J1261 to R. Port and by the National Institute of Mental Health, grant number MH10667 to J. Devin McAuley.

NOTES

1. There are other difficulties with a "neural spectrogram" model of auditory memory. The main one is, how could it be implemented? The memory could not plausibly be sampled in time, since this should lead to obvious aliasing artifacts for inputs at certain frequencies. Nor could it just be an exponentially decaying trace for independent frequency bands of the spectrum (like the way a piano "records" your voice if you shout at it after lifting the dampers with the pedal). If it worked like this, then later sounds in a given frequency range would tend to be confused with earlier sounds, which does not seem to be the case. Of course, it is undeniable that we *do* have introspective access to recent sound, at least when the sounds are familiar. For these, we probably store some kind of descriptive labels. Evidence for the necessity of learning complex patterns comes from research on patterns that are novel but complex. It is known that subjects cannot make good discriminations of complex patterns that are unfamiliar (Espinoza-Varas and Watson, 1986; Spiegel and Watson, 1981; Port, 1990).

2. By "weaker" and "stronger" measures of time, we refer informally to the set of invariance transformations that are permitted on the scale (Stevens, 1951; Port, 1986), i.e., the transformations that do not disturb the temporal description. For serial order, many complex transformations on the duration of component events are possible without disturbing serial order. For phase angle measurement, only durational transformations that preserve relative duration are allowable. Absolute measurements permit no durational changes at all.

3. These are actually schematic diagrams that illustrate the principles at work. When our simulations (Anderson and Port, 1990; Cummins, 1993) were carried out, the state space was of higher dimension (typically around 12) and the set of targets and distractors was considerably larger (as large as ten each).

4. Of course, the primitive dynamic memory described here can only track one familiar sequence at a time. One would need several distinct simple dynamic memories to track several patterns simultaneously. Presumably animal auditory systems can deal with this for at least several overlapping patterns.

REFERENCES

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

Anderson, S., and Port, R. (1990). *A network model of auditory pattern recognition.* (Technical report no. 11.) Bloomington: Indiana University Cognitive Science Program.

Anderson, S., and Port, R. (1994). Evidence for syllable structure, stress and juncture from segmental durations. *Journal of Phonetics, 22,* 184–217.

Anderson, S. E. (1994). *A computational model of auditory pattern recognition.* (Technical report no. 112.) Bloomington: Cognitive Science Program, Indiana University.

Baddeley, A. (1992). Working memory. *Science, 255,* 556–559.

Bernstein, N. (1967). *The coordination and regulation of movements.* London: Pergamon.

Bregman, A. S. (1990) *Auditory scene analysis: the perceptual organization of sound.* Cambridge, MA: MIT Press.

Browman, C., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201–251.

Chomsky, N., and Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Crowder, R., and Morton, J. (1969). Precategorical acoustic storage. *Perception and Psychophysics, 5,* 365–373.

Cummins, F. (1993). Representation of temporal patterns in recurrent neural networks. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 377–382). Hillsdale, NJ: Erlbaum.

Das, S., Giles, C. L., and Zheng Sun, G. (1992). Learning context-free grammars: capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 791–796). Hillsdale, NJ: Erlbaum.

Elman, J., and Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America, 83,* 1615–1626.

Espinoza-Varas, B., and Watson, C. (1986). Temporal discrimination for single components of nonspeech auditory patterns. *Journal of the Acoustical Society of America, 80,* 1685–1694.

Fowler, C., Rubin, P., Remez, R., et al, (1981). Implications for speech production of a general theory of action. In B. Butterworth, (Ed.), *Language production* (pp. 373–420). New York: Academic Press.

Fraisse, P. (1957). *Psychologie du temps.* Paris: Presses Universitaires de France.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: speech language, and motor control. In E. Schwab and H. Nusbaum (Eds.), *Pattern recognition by humans and machines: speech perception.* Orlando, FL: Academic Press.

Grossberg, S. (1988). *Neural networks and natural intelligence. Cambridge, MA:* MIT Press.

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to speech analysis: the distinctive features and their correlates.* Cambridge, MA: MIT Press.

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

Kelso, J. S., Saltzman, E., and Tuller, B. (1986). The dynamical perspective in speech production: data and theory. *Journal of Phonetics, 14,* 29–59.

Klatt, D. H. (1976). The linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208–1221.

Klatt, D. H. (1977). Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America, 62,* 1345–1366.

Klatt, D. H. (1980). Speech perception: a model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech,* (pp. 243–288). Hillsdale, NJ: Erlbaum.

Kolen, J. F. (1994). Fool's gold: extracting finite state machines from recurrent network dynamics. In *Advances in Neural Information Processing Systems, 6,* in press.

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.

Large, E. W., and Kolen, J. F. (1994). Resonance and the perception of musical meter. *Connection Science,* in press.

Lee, C. H., Rabiner, L. R., and Pieraccini, R. (1992). Speaker independent continuous speech recognition using continuous density hidden Markov models. In *Speech recognition and understanding. Recent advances, trends and applications* (pp. 135–163). *Proceedings of the NATO Advanced Study Institute, Cetraro, Italy, July 1–13, 1990,* Berlin: Springer Verlag.

Lee, K. (1992). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. In *Speech recognition and understanding. Recent advances, trends and applications (p. 133). Proceedings of the NATO Advanced Study Institute, Cetraro, Italy, July 1–13, 1990,* Berlin: Springer Verlag.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Lesser, V. R., Fennel, R. D., Erman, L. D., et al. (1975). Organization of the Hearsay-II speech understanding system. *International Conference on Acoustics, Speech, and Signal Processing, 23,* 11–23.

Lisker, L., and Abramson, A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20,* 384–422.

Lisker, L., and Abramson, A. (1971). Distinctive features and laryngeal control. *Language, 44,* 767–785.

Massaro, D. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79,* 124–145.

Massaro, D. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

McAuley, J. D. (1994a). Finding metrical structure in time. In Mozer, M. C., Smolensky, P., Touretzky, D. S., et al. *Proceedings of the 1993 Connectionist Models Summer School* (pp. 219–227). Hillsdale, NJ: Erlbaum.

McAuley, J. D. (1994b). Time as phase: A dynamic model of time perception. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society* (pp. 607–612). Hillsdale, NJ: Erlbaum.

Michon, J. A., and Jackson, J. L. (1985). *Time, mind and behavior.* Berlin: Springer Verlag.

Neisser, U. (1967). *Cognitive psychology.* New York: Appelton-Century-Crofts.

Pollack, J. B. (1991). The induction of dynamical recognizers. *Machine Learning, 7,* 123–148.

Port, R. (1986). Invariance in phonetics. In J. Perkell and D. Klatt (Eds.), *Invariance and variability in speech processes.* Hillsdale, NJ: Erlbaum.

Port, R., and Dalby, J. (1982). C/V ratio as a cue for voicing in English. *Journal of the Acoustical Society of America, 69,* 262–1274.

Port, R., Reilly, W., and Maki, D. (1987). Use of syllable-scale timing to discriminate words. *Journal of the Acoustical Society of America, 83,* 265–273.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America, 69,* 262–274.

Port, R. F. (1990). Representation and recognition of temporal patterns. *Connection Science, 2,* 151–176.

Port, R. F., and Cummins, F. (1992). The English voicing contrast as velocity perturbation. In *Proceedings of the Conference on Spoken Language Processing* (pp. 1311–1314). Edmonton, Alberta: University of Alberta.

Port, R. F., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America, 81,* 1574–1585.

Povel, D., and Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2,* 411–440.

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

Saltzman, E., and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382.

Shamma, S. A. (1989). Stereausis: binaural processing without neural delays. *Journal of the Acoustical Society of America, 86,* 989–1006.

Sorkin, R. (1987). Temporal factors in the discrimination of tonal sequences. *Journal of the Acoustical Society of America, 82,* 1218–1226.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74,* 1–29.

Spiegel, M. F., and Watson, C. S. (1981). Factors in the discrimination of tonal patterns. III. Frequency discrimination with components of well-learned patterns. *Journal of the Acoustical Society of America, 69,* 223–230.

Stevens, K. N., and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358–1368.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens, (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Tajima, K., Port, R., and Dalby, J. (1993). Influence of timing on intelligibility of foreign-accented English. *Journal of the Acoustical Society of America, 95,* 3009.

Torras, C. (1985). *Temporal-pattern learning in neural models.* Berlin: Springer-Verlag.

Van Gelder, T. and Port, R. (1994). Beyond symbolic: toward a *kama-sutra* of compositionality. In V. Honavar and L. Uhr (Eds.), *Artificial intelligence and neural networks: steps toward principled integration* (pp. 107–125). New York: Academic Press.

Warren, R. M. (1993). Perception of acoustic sequences: global integration versus temporal resolution. In S. McAdams and E. Bigand (Eds.), *Thinking in sound: the cognitive psychology of human audition,* (pp. 37–68). Oxford University Press, Oxford.

Watrous, R. L. (1990). Phoneme discrimination using connectionist networks. *Journal of the Acoustical Society of America, 87,* 1753–1772.

Watson, C., and Foyle, D. (1985). Central factors in the discrimination and identification of complex sounds. *Journal of the Acoustical Society of America, 78,* 375–380.

Williams, R., and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1,* 270–280.

Guide to Further Reading

The problem of timing in speech was considered in Port (1986, 1990). Both papers were partially inspired by the classic work on timing in behavior by Lashley (1951) and on measurement scales by S. S. Stevens (1951). Dealing with time in connectionist networks has been addressed in a simple way by Elman (1990) and in Lippman's review (1989) of neural networks for speech recognition. Many of the basic phenomena of speech timing, especially in English, were reviewed in Klatt (1976). One important recent model of sequence recognition in neural networks is Sven Anderson's thesis (1994).

Anderson, S. (1994). *A computational model of auditory pattern recognition.* (Technical report no. 112). Bloomington: Indiana University Cognitive Science Program.

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

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of Acoustical Society of America, 59,* 1208–1221.

Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.

Lippman, R. (1989). Review of neural networks for speech recognition. *Neural Computation 1*, 1–38.

Port, R. (1986). Invariance in phonetics. In J. Perkell and D. Klatt (Eds.), *Invariance and variability in speech processes 1* (pp. 540–558). Hillsdale, NJ: Erlbaum.

Port, R. (1990). Representation and recognition of temporal patterns. *Connection Science 2*, 151–176.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S. (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.