TECHNICAL REPORT NO. 310

Representation and Recognition of Temporal Patterns

by

Robert Port

June 1990

COMPUTER SCIENCE DEPARTMENT
INDIANA UNIVERSITY

Bloomington, Indiana 47405-4101

Representation and Recognition of Temporal Patterns

Robert Port*Department of Linguistics,
Department of Computer Science
Indiana University, Bloomington, Indiana 47405.

Abstract

How can a nervous system represent for itself the temporal relations of patterns that it knows? In order to label auditory patterns, the nervous system must store early portions in order to identify the whole. Both linguists and engineer-scientists have a similar need to record spoken words. This paper reviews 3 basic models for handling the information-collection problem that supports pattern recognition whether by scientists or others. Many of these techniques have been implemented in connectionist networks. In linguistic models for words, there are only ordered symbols, that is, either phonemic segments or words. In engineering and speech science, time windows are built that store the entire signal and allow parametric description of time. But such windows are not plausible for nervous systems. A third alternative is a memory in the form of a dynamic system. These models are driven through a trajectory in state space by the input signals. Thus, the recognition process for familiar patterns produces a distinct trajectory through state space for each learned pattern. Among the advantages of such a system is that (1) it tends to recognise patterns despite changes in the rate of presentation, and (2) the system can be run continuously yet will respond as quickly as possible at appropriate times. Evidence will be reviewed about human auditory memory for complex tone sequences. The data suggest that human auditory memory exhibits many similarities to the dynamic model.

1 Introduction

How can a nervous system represent for itself the temporal relations of patterns that it knows? What happens when you recognize an auditory pattern – like a word, a dog bark, a train of footsteps or a familiar dance rhythm? To reach a state of recognition,

*I am grateful to Sven Anderson for many contributions to the research described here. I am also grateful to Charles Watson, Gary Kidd, Michael Gasser, Jungyul Suh and John W. R. Merrill for helpful discussion of these ideas. This research was supported in part by the Air Force Office of Scientific Research, Grant 870089, and by the National Science Foundation, Grants DCR-8505635 and DCR-8518725. To appear in *Connection Science*, 1990.

some representation for the stimulus over time must have been created that is sufficient to permit recognition to occur. The basic issue is how can complex patterns in time be recognized?

Despite Lashley's proclamation [Lashley, 1951] that temporal patterns are central to the study of cognition, the fundamental issues surrounding the representation of time have not attracted a great deal of attention over the decades. In contrast to Lashley's primary interest in motor control, our concern here is primarily with the problem of perception of patterns in time. It turns out that some of Lashley's theoretical suggestions sound quite modern and are applicable to modern thinking about perceptual models.

This paper reviews three general schemes for the representation of sound as events in time. Although many examples are described, the primary emphasis is on connectionist models, since they seem the most promising general framework for cognitive models. These familiar frameworks serve as a primitive typology of models. As far as I can tell, they span the range of basic theories addressing this problem. In the next section, evidence from experimental psychology is described that helps clarify the problem further. Research on the perception of complex tone patterns – sequences of up to 10 randomly chosen tones – suggests that one of the three representational frameworks is better, since the model exhibits several properties exhibited by human subjects when they learn to recognize complex sound patterns.

The issues here are fundamentally part of a general theory of perception, but I will focus on techniques for representation of events distributed in time that support perceptual recognition. The discussion will also be directed toward issues in speech and language. Although there are a great many theories about perception, there are really only a couple of theories about how temporally distributed events might be represented either in natural cognition or in models of cognitive-like processes.

The first model for temporal representation to be discussed below is the Linguistic View of the symbolic structure of cognition. The well-known Jakobsonian manifestation of this general lingiustic model will be illustrated. In classical linguistics, words were said to be spelled as ordered strings of segmental phonemes – consonants and vowels. Incidentally, S. S. Stevens' typology of psychological scales as *nominal, interval, integer* and *ratio* [Stevens, 1951] provides helpful terminology for discussions of time. Clearly both phonemes and words, if they are part of the perception of speech, comprise an ordinal scale. For example, like the symbols of a phonetic transcription or letters on a page, their spacing and size could be altered without changing their meaning, and the number of symbols used to transcribe an utterance does not provide a reliable measure of duration in real time.

Two analogs of real world temporal order can be defined in relation to the symbolic model. First, within the symbol-system itself, locations in a string or buffer precede and follow one another. On the other hand, in the external controller that 'runs' the symbol system (that is, the operating system, programmer or whatever), the states of a derivation precede and follow one another in the non-historical, timeless time in which control operations for mathematical models are carried out.[1]

[1]The treatment of program states as time sounds absurd in artificial intelligence. No one would

The linguistic answer to the Lashley's problem was proposed by Chomsky. It is that we should handle syntactic and phonological patterns in time by means of strings of symbols. Indeed all current linguistic theory, whether in phonetics, phonology or syntax, looks for the effects of rule-governed symbol string manipulation. This presumes a discipline of phonetics that will supply to linguistic theory a set of discrete symbolic units [Ladefoged, 1989].

The second framework for temporal pattern specification is the Engineering Approach. Those with practical concern for real sounds, like phoneticians and engineers, needed mechanical techniques to store and process the sounds. Time is transformed into an analog signal or a vector of numbers that can be recorded mechanically and manipulated in useful ways. Obviously, there are many techniques that fall under this heading. Simply making a sound spectrogram uses a clocklike device to translate temporal patterns into spatial patterns. And speech recognition systems make a 'window' on the signal and interpret spatial position in the symbol string to infer time. By such techniques, an unlimited amount of information about events in time can, in principle, be extracted and represented as numbers. A vast variety of quantitative relations at various levels of abstraction can be defined, like the ratio of two durations. But, what is frequently ignored is that there must be two systems here: (1) a program, the statement of processes and relationships, and (2) a controller, like the operating system, that reads through the statements of the program, interpreting and executing them.

The third framework, and the most subtle to design, is a Dynamic Model. Such a system learns to respond dynamically to familiar patterns in time. This kind of model reaches and remains for a while in a unique state for each trained pattern. It reaches this stable state at the end of a distinctive trajectory of states. The representation of time here could be said to be either the distinct trajectory (that is, a dynamic pattern) or some terminal stable state itself. This class of systems is not yet well-developed, but they are important since they have the advantage of not requiring an external control system. Decisions about what to do next are built in to the dynamics of representations.

These three models or frameworks embody different assumptions about what kind of temporal information the central nervous system can use and how this information is to be extracted. I will argue that there are actually strong empirical reasons for continuing to explore the dynamic model. The first matter that needs to be addressed, however, is the question of what a temporal pattern is. This is attempted in the next section.

The Problem of Temporal Patterns. What kind of temporal patterns are relevant to the study of cognition? All sound consists of variations in air pressure over time. But not all sound is cognitively temporal. On the other hand, some patterns can be quite long even though we may know every detail. The fundamental issue is what properties of sound

---

imagine that the order of states in a program implies any cognitive claims. Within linguistics, however, there have been debates about rule ordering which assume that the order of derivational states somehow resembles real time. Some states are reached 'earlier' and others 'later'. Recently, new appoaches to phonology inspired by connectionism have attempted to have rules apply simultaneously ([Lakoff, 1988,Hare, 1990,Gasser and Lee, 1989] or partly simultaneously [Wheeler and Touretzky, 1989].

patterns are specified by time and what by place. There is a frequency representation in the auditory nerve that is extracted from ambient sound by the ear. From the highest audible frequencies down to a low range (perhaps as low as 100 Hz) there are fibers that fire most strongly for particular frequencies. And fibers representing nearby frequencies are physiological neighbors. This is effectively a place code, or *tonotopic representation*, for frequencies. We may think of the sound as filtered through critical-band filters with a range of integration times (proportional to the wave length). At the same time, however, the auditory nerve also exhibits strong time-synchronous activity: cells tend to fire whenever the input wave reaches a particular phase angle [Sachs and Young, 1980,Moore, 1982]. By means time-synchronous firing, there is a direct temporal code not only of very slow events, but even of the successive 'clicks' on each cycle of tones as high as 4 kHz ([Moore, 1982]). It is obvious that very long-period events (like a ringing telephone) must be processed as events in time. But the brain also has the information available to process many very fast events as temporal – even though a place code is also produced by the ear. Only a little is known about the role of these overlapping sources of frequency information (see [Warren and Bashford, 1981]).

For the purposes of computational modelling of hearing and perception, it is convenient to settle upon some single compromise integration window and sampling rate (see [O'Shaughnessy, 1987] for general discussion). There many variants specialized for various purposes. Typical integration windows for speech signals lie in the range of 2 ms (like a wideband spectrogram) to 20 ms (like a narrowband spectrogram) [Klatt, 1986],[Carlson and Granstr
For modelling purposes, it probably does not matter much how these details are implemented as long as both frequency (that is, place) information and temporal information are available. Our concern in this paper is with how to use the information that does *not* have a natural spatial representation.

2 Temporal Representations

In the following sections, three frameworks for representing temporal patterns in a perceptual or recognition system are sketched. These models have been popular within different disciplines dealing with speech since they seem to address the necessary issues. First, we should consider some apriori constraints on the representation of signals to support perception of temporal patterns. Very long and abstract patterns in time – such as a familiar melody – should probably be represented in abstract units like the notes of a symbolic musical scale (since a melody should be invariant over changes in key and musical instrument, for example). But problematic issues arise for short duration patterns, like stop consonants and syllables – not to mention the sounds of chair squeeks, light switches, breaking glass and refrigerator doors. These patterns are short enough so that there are no obvious 'symbol-sized' pieces from which they could be represented. Until some initial code for complex temporal patterns can be modelled, we are unlikely to succeed in representing higher-level units – like musical notes, syllables and words. Many examples of speech perception show that listeners make use of detailed temporal prop-

erties of words as distributed over a range of a syllable or more. Changes in temporal detail can sometimes change one word into another. Such results suggest that the temporal invariance of words as they are actually pronounced is not nearly so abstract as an ordinal scale. The kind of timing distortions that are permitted by an ordinal scale for time go far beyond the kind of rate change transformations that one observes in speech (see [Dorman et al., 1979,Port, 1986,Port and Dalby, 1982]. Observations such as these make one skeptical about current linguistic models of memory for the phonetic and phonological form of words.

2.1 Linguistic Viewpoint

In a strictly symbolic model like theoretical linguistics, there is no concept of time other than in ordering symbols. In linguistic models of phonology and syntax, the sequential property of words – that is, their occurrence in historical or 'real' time – is expressed by concatenating symbols in a particular order. The symbols are phonetic or phonological segments, such as those of the IPA (International Phonetic Association), Jakobson [Jakobson et al., 1952] or Chomsky & Halle's *Sound Pattern of English* ([Chomsky and Halle, 1968]). Although a string of phonemes may look like an integer scale (like 'beads on a string' as they used to say in the 1940's), linguists actually rely on ordinal 'distance' measures (like 'A and B are adjacent/nonadjacent'). Symbols can be defined that are intended to be interpreted as meaning 'longer' and 'shorter', but this doesn't count as having to do with time (see [Lisker and Abramson, 1971]). Since acoustic inputs are actually distributed in real time, a perceptual system must have some sort of 'front end' that produces discrete segmental units from this continuum. The development of this front end is normally taken to be the task of experimental phonetics [Chomsky and Halle, 1968]. As shown below, the feature detectors at the front end are primarily conceptualized as acoustic filters that integrate over some fixed time window on the signal. Indeed, it is impossible to imagine a perceptual system that does not integrate events when they are close enough together. This integration process necessarily results in loss of information about the precise location of events within the integration window. The empirical issue is whether that hypothesized loss in the case of particular features is also observed in human listeners.

Jakobson's Distinctive Features Since Jakobson's early work, feature-detector systems have been conceptualized as devices that integrate information over a certain time window ([Jakobson et al., 1952] [Fant, 1973]). Thus, the feature *acute* was defined (in [Jakobson et al., 1952]) as a spectrum that has relatively more energy above 4 kHz than below 4kHz (assuming other parameters are held constant, or *ceteris paribus*, as Jakobson put it). It was hoped that for each feature, some filter could be defined which would indicate when a particular linguistic feature occurred in the signal. If static detectors that simply integrate their inputs are sufficient to permit identification of these features, then a complete theory of speech perception might be constructed that would never have to deal with time as a scalar parameter (nor general spectral parameters either, apparently). Such a result would be

very attractive to those who attempt to demonstrate that serial order is the only temporal relationship required for human cognition. If time-integrating features like this can be defined effectively, then a linguistic model could be proposed that jumps in one step from real-time integration of the auditory signal (that is, averaging over time) to a serially ordered symbolic structure for language and other higher cognition.

This perceptual model was conceptualized as a bank of feature detectors that examine the incoming acoustic signal. The detectors would fire synchronously once for each segment. In this way, not only is the spectral space coded into a smaller set of properties, but time has also been converted to a string of time-integrated objects. Jakobson *et al.* were unclear how the correct center point for each segment was to be found as the signal streamed through the detectors. In a Pandaemonium-like [Selfridge, 1959] or connectionist implementation of such features, each feature could inhibit its competitors. Thus there would only be one winner from among competing sets at a time.

Most of the 12 features in the system of Jakobson, Fant and Halle are defined in terms of information integrated over time windows of roughly 20-50 ms. Thus, the features [grave], [acute] and [compact] were defined in terms of a spectrum with, respectively, downard tilt (that is, more energy in lower frequencies), upward tilt, and band-passed with a center-frequency at around 3000 Hz. For stops, this window was to be centered over the stop burst. Depending on the relative output of two opposed filters (for each $\pm$ value), the stop would be categorized as plus or minus for each feature. Thus, the linguistic features from which phonemes were defined were to be directly extracted from the signal by using a set of integration frames. A few features like [interrupted] were defined by a change in value between neighboring temporal slices (eg, from low-amplitude to high-amplitude).

Although the Jakobson-Fant-Halle system could never actually be implemented as a speech recognition device,[2] the conceptual model continues to thrive within linguistics and linguistic phonetics despite experimental phonetic evidence of the importance of the scalar properties of speech timing. For example, Stevens and Blumstein explored variants of the Jakobsonian place features by using fixed integration windows for carefully constructed static spectral templates [Stevens and Blumstein, 1981] [Stevens, 1983]. Other work showed that performance would improve if dynamic properties of the spectra were considered [Kewley-Port, 1983].

Incidentally, recent developments in phonology [Goldsmith, 1976,Clements, 1985] have expanded the traditional model of segments by incorporating 'autosegmental' data structures. These allow each articulatory subsystem (or 'tier') to define its own sequence relations. And all autosegmental tiers (such as the lips, nasality, tone, etc) are linked together in a single spatio-temporal time line. This model seems attractive since it no longer forces

[2]Despite all the attractive properties of Jakobson's vision, the model was flawed because of the *ceteris paribus* clause. This phrase, almost a refrain in the original technical report, meant that the distinctive features were always defined by holding everything else constant while the comparison was made. That is, attention was always focussed on *minimal pairs*. But discrimination between pairs of competitors is not what speech recognition or speech perception requires. To be successful, it must *identify something* – anything that can be found. Identification means distinguishing some feature of the sound from *all other features of sound*. This is vastly more difficult that distinguishing minimally different word pairs.

linguistic time onto a discrete scale (like orthographic letters). Instead, the ordinal nature of linguistic elements is clarified and enriched. [3]

There has long been evidence that many temporal properties of speech play critical roles in the production and perception of speech (see [Lehiste, 1970,Klatt, 1976] for reviews, or [Port, 1981,Port and Rotunno, 1979,Port and Crawford, 1989]). Numerous disputes have arisen over the years as to whether information about scalar time, measured in rational numbers, rather than simply timeless feature states need to be incorporated into theories of speech perception and production.[4]

2.2 Engineering Solutions

Over the past 40 years, researchers with both scientific and engineering interests have developed a variety of practical means of representing speech signals. These approaches have ignored the traditional linguistic view of ordinal segments – segments that can only be obtained from a linguist or phonetician. The engineering tradition is represented by a huge body of research in many disciplines. A sampling of examples of programs and models that process temporal information are mentioned below.

One of the earliest and most long-lived devices that allowed scientists and engineers to study speech sounds was the analog sound spectrograph that became available around 1950.

Sound Spectrogram Phoneticians need tools for the study of speech sounds, of course, so they make sound spectrograms on sheets of paper: a display of frequency by time with intensity displayed as darkness. Like tape recorders, clocks and digital-analog converters, the sound spectrograph uses constant motion to translate time into distance. Much interesting research has been conducted on spectrograms and their digital successors. The sound spectrogram in Figure 1 below has a distance measure for time. The figure shows a pair of English words and demonstrate the kind of information available in spectrograms. Speech scientists take temporal measurements from such displays and infer characteristics of speech production and speech perception from them (see, for example, [Lisker and Abramson, 1964,Klatt, 1976]).

[3] Autosegmental models clarify the original insight of segmental descriptions of words since one is no longer tempted to ask 'How many segments away from phoneme A is phoneme B?', as though they were beads. I doubt any linguist ever wrote a rule involving counting segments. It may make sense to speak of "4-letter words" but linguists don't ask about 4-*phoneme* words – because linguists think of phonemes as ordinal.

[4] Another prominent example of such a dispute between phoneticians and supporters of ordinal time is the issue of voice-onset time as a cue for the feature [voice]. Lisker and Abramson [Lisker and Abramson, 1964] proposed that voice-onset time, the interval between the burst on an utterance-initial stop and the voicing onset (measured as a scalar) is an important cue for speech perception in English. Their results were reinterpreted, however, in terms of timeless distinctive features by [Chomsky and Halle, 1968] and [Halle and Stevens, 1980]. Lisker and Abramson responded by insisting that scalar information about temporal structure cannot be ignored [Lisker and Abramson, 1971]. The issue was never resolved since it involved fundamental assumptions about models for auditory memory for speech.

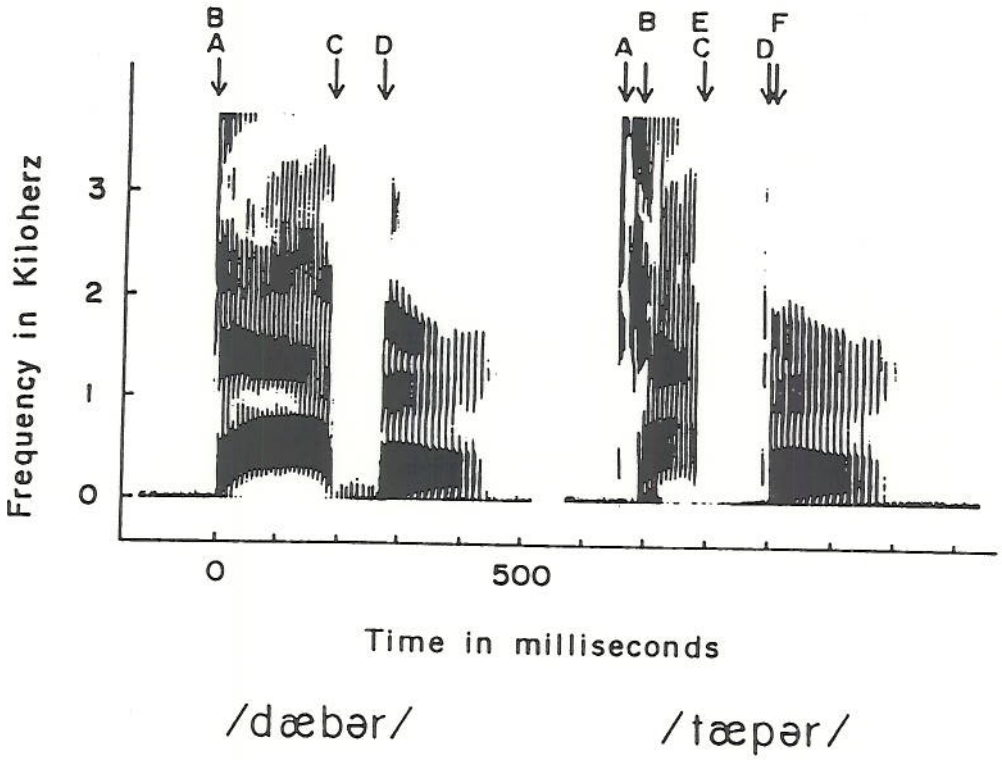

Figure 1: Spectrograms of the words *dabber* and *tapper* showing A, the point of the stop burst, B, the point of the voicing onset, C, the onset of closure for the medial stop, D, release of the medial stop, E, voicing offset for the medial stop, and, F, voicing onset for the medial stop. It can be seen that the two segmental differences between the words ([d] vs. [t] and [b] vs. [p]) are manifested as several durational differences: The first vowel is longer before the [b] than the [p], and conversely, the [b] closure is shorter than the [p]. But *tappers* offers several possible locations of the 'vowel onset'. But, in any case, how is vowel onset defined also for *flubber* and *rubber*? Manipulation of these parameters in synthetic speech stimuli shows that they are used by listeners to identify words.

Although the notion of 'echoic memory' has been proposed [Neisser, 1967] [Crowder and Morton, 1 evidence that this representation is the raw unprocessed signal is based on stimuli that are either very simple (like tones) or very familiar (like speech). On the other hand, when two familiar auditory patterns, such as one speech and the other nonspeech, are presented simultaneously. Listeners tend to segregate (or parse) them into independent streams that correspond, in many cases, to familiar patterns ([Bregman and Campbell, 1971] or, for a review, see [Handel, 1989]). The temporal alignment of the unrelated patterns with respect to each other (like speech with a superimposed click) is only poorly represented [Bever, 1973]. Such a result would not be expected if a representation of the raw signal itself were available to subjects.

On the other hand, we can collect evidence about the kind of information that perceptual systems must be using when listening to speech. It can be shown that many of the temporal details of natural speech are detected and used by listeners (see [Klatt, 1976,Repp, 1984] for reviews). By and large, most prominent temporal regularities can be shown to play a role in speech perception if a sensitive experimental task is designed [Liberman et al., 1967][Dorman et

Dynamic Time Warping Despite evidence of the role of temporal patterns in the specification of words (at least from experiments on minimal pairs), this information could not be incorporated in speech recognition very easily (see [Waibel, 1986]). One reason is that easily intelligible tokens of real words are normally produced with a vast variety of temporal patterns. This variation is due to speaking tempo in part, but also due to many other factors. Speech recognition techniques made practical progress in the 1970s (as reviewed in [Lea, 1980] and [Vassiere, 1985]) by ignoring linguistic segmental descriptions. One result was the dynamic time-warping (DTW) algorithm as a way to get rid of spurious temporal information in isolated-word recognizers. DTW permits the comparison of a test item with templates in such a way that errors due to differences in the timing patterns are minimized [Itakura, 1975,Sankoff and Kruskal, 1983]. The algorithm allows comparisons to be based almost entirely on serial order.[5] After collection and averaging of a set of 'training tokens' for items in the vocabulary, the recognition system constructs a template for each word against which new test tokens can be compared. The dimensions of the template are discrete time and frequency (or other parameters based on data input). However, since actually occurring input items occur at a wide variety of rates and styles, time must be normalized. A technique that simply scales all durations by a constant amount does not work well for several reasons. First, rate variation produced by human speakers is much more complex and nonlinear ([Klatt, 1976,Port, 1981]. Secondly, it is difficult to normalize by, say, 10% when the spectral frame sizes are fixed. Both problems mean that reducing error in one temporal region will tend to increase it at the other end of the template. So, dynamic time warping, as shown in Figure 2, is done by finding the monotonic mapping from the template vector to the test item vector that produces minimum error (where error is the difference between the test vector and the reference vector) across the entire mapping. The vector for each time slice is subtracted from all the neighboring vectors in the template. The algorithm finds a path with the lowest total error through that matrix starting at (1,1) and ending at (I, J). This approach is motivated by a strong assumption about speech: that speech production can change tempo so rapidly that it could be changed by a large amount between each time frame.

Although dynamic time-warping was never intended as a cognitive model, it neverthless offers a provocative a way of thinking about time. It assumes, just as the linguists insist, that linguistic information lies only in *the sequential order of the states* (although, of course, constraints on the warping path are always applied in practice). Of course, the states here are spectral slices, not segmental symbols as in linguistics. In acknowledging that the states are distributed in highly irregular ways, it treats all temporal structure as a kind of noise to be ignored in this clever way.

Even though reassuring in some sense, to the traditional linguistic view, there is much evidence that scalar values of time are more important than that. There is too much

[5]The mapping between template and test item makes only very weak assumptions about contraints on timing variation. As noted above, there is information in many details of speech timing that could be used to help recognize words. A primitive means to do this was demonstrated in [Robert Port and Maki, 1988] even though these authors did not claim that their technique would be generally practical.

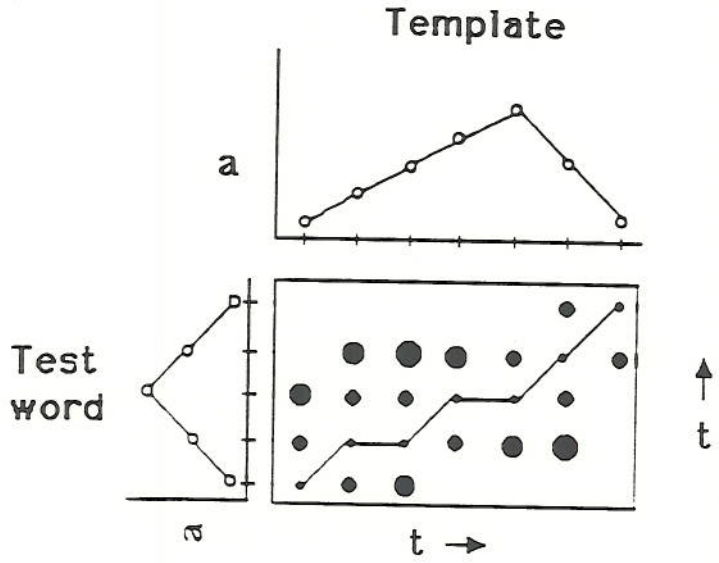

Figure 2: Dynamic time warping is done by finding a minimum error mapping from the reference (or template) vector to the test item vector. The vector for each time slice is subtracted from all other vectors. The magnitude of those differences is displayed here as circles. The algorithm finds a path through the matrix starting at (1,1) and ending at (I, J), where there are I time steps in the template and J steps in the test item. It accumulates the error across all the $max(i, j)$ pairings. This error is used for comparing different words. The word with the smallest error after time warping that meets some threshold is chosen.

phonetic evidence that subtle details of speech timing carries usable information about word identities [Klatt, 1976,Port, 1986]. Engineers and phoneticians have preferred to act as though listeners have available to them a time window on the signal, a stretch of the signal that is available at one time, like a 'neural spectrogram'.

Nonoverlapping Windows As in conventional approaches to speech recognition, many systems within the connectionist school employ windows that are long and more or less nonoverlapping. These present the network with the entire bandwidth of a long enough stretch of signal to allow the whole target pattern to be displayed. The basic idea is to have a separate set of nodes to represent raw inputs from each time slice. Thus, these models typically have a single output category node for the whole window.

One prominent example among network models for a speech recognition task is the model of Elman & Zipser (1988, section 3) . This feedforward system (trained with backpropagation) was presented with speech samples in a simultaneous window of 64 ms in 20 slices. For each slice, 16 normalized spectral energy measures were provided. The speech samples were many productions of the 9 syllables: /bi,ba,bu, di,da,du, gi,ga,gu/ by a single speaker. Thus the network had 320 inputs, only 3 hidden nodes and 9 output nodes and was trained to label the spectrum in the window. Elman and Zipser showed that this sort

of network successfully categorizes stop-vowel syllables despite variation across hundreds of tokens. At first glance, the network appears to capture phonetically useful information about location in time in terms of a spatial representation within the window. But from the standpoint of the network, the nodes representing each time slice are just parallel input channels. They do not have even intrinsic serial order. Thus they actually constitute only a *nominal scale* for time [Stevens, 1951]. The reason that the system can differentiate the stops from each other and the vowels from each other is that the stop burst always occurs at about the same place in each token (so the 'shift-invariance problem' is minimal). Clearly the window duration must be fixed since window size determines network size.

As a theory of perception of temporal patterns, the nonoverlapping window solution, with or without labels indicating serial position within the window, has little to support it. If one tried to run it continuously, there would be strobe-like effects due to interference with the periodic 'blanking interval' – the time point where window contents are switched. Inputs that are periodic at the frequency of buffer switching or its multiples should be susceptible to auditory disappearance. Since no information about the signal is obtained prior to filling the window, this window would have to have a fixed duration (although a stochastically varied window duration could help avoid stroboscopic interference). Only after the window is full, can long patterns be used to influence processing itself.

Of course, phoneticians have certainly never claimed that the $f \times I \times t$ spectrogram is a part of the theory of auditory perception, but reflection shows that something like it must implicitly be postulated if parametric measurement of time intervals is to be available to human listeners. (Of course, humans and animals are known to be poor at parametric measurement of time even though a 'chronotopic' window should make that easy.)

If time is converted into place, such models explicitly separate the representation of time from perceptual analysis of time. But parsing the input this way requires breaking experience of the world up into temporal chunks, that imply periodic alternations of information collection and information processing. The motivation for this kind of 'time multiplexing' is that the window must be long in order to recognize long patterns. In order to be long enough for the longer patterns, the ability to respond to short patterns is artificially delayed – even though some short patterns require rapid reaction time. Surely, the criterion of a quick response is important enough that long patterns must be extracted by a technique that permits continuous streaming of inputs and response that is as early as the information in the pattern permits.

Although the static window models mentioned so far have windows that have little overlap (since it is assumed that each section needs to be analyzed only once), it is also possible to have reduced amounts of overlap up to the point at which the window is as narrow as possible, and it slides over the spectral frames with each increment of time.

Iterative Nets with Delay Lines One way to move toward continuous behavior without having recurrent edges (since backpropagation works only for feedforward architectures) is to unwrap the network in time. Thus a feedforward net for processing long patterns is analogous to a much smaller network with recurrent edges, as illustrated in Figure 3.

This variant of the time window has inputs that arrive at different places in the net [Hinton, 1988,Elman and McClelland, 1986,Lang et al., 1990]. At each time slice after the first, what is retained about earlier events is an analysis of those events, not the original signal itself. But these have several problems that limit their biological plausibility. First, at the close of stimulus presentation, they must be reset to an initial activation state for presentation of the next stimulus. Second, they have a set of inputs and outputs for each iteration of the network (that is, for each time frame of possible patterns), rather than one set of outputs for each learned pattern. Thus, some additional technique must be used either to obtain a single categorical response to the entire pattern.

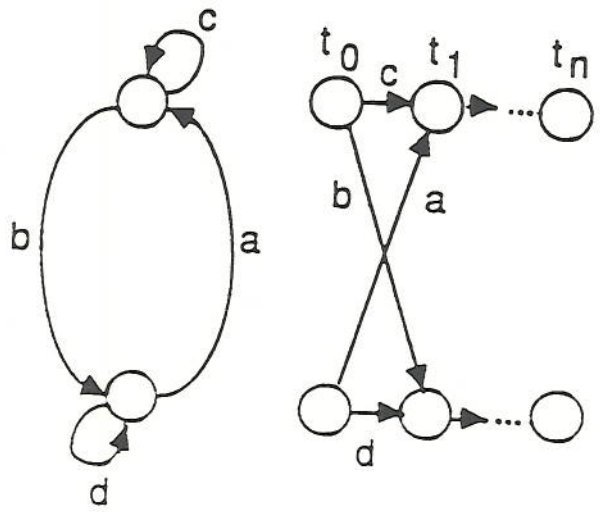

Figure 3: A simple iterative network in which stimulation for different time frames are supplied to different input nodes. A single kernel network at the left is basically repeated for each time slice. The flow of activations through the network simulates recurrent activity through time. After each trial, the system resets all activations and the next window begins again at t_0. One advantage is the architecture remains feedforward so that backpropagation can be used.

The familiar NET-talk system [Sejnowski and Rosenberg, 1986] for translating orthographically spelled text into phonemic transcriptions could be streamed this way. The system takes 7 input characters (ordinary letters, spaces and punctuation), using 8 1-bit input nodes per character, and outputs a single phonemic character corresponding roughly to the middle input character. Then it slides over by one symbol of input and repeats the operation. In effect then, each node in the hidden layer is summing inputs from 3 symbols forward (relative to the output character) and 3 backward across the input. These connections across time are also sometimes called *delay lines.*

In a recent study, Lang, Waibel and Hinton [Lang et al., 1990] demonstrate many interesting properties of models that collect information from continuously streamed input. There are many other studies as well [Tank and Hopfield, 1987,Elman and Zipser, 1988] and [Waibel et al., 1988,Watrous, 1990]. In such a system, events at several adjacent points in time are presented simultaneously to a single node in the network. It is like

a time window since physical places (that is, input lines) represent distinct points in time. But in Lang et al's model, the network itself looks only at a short stretch of signal. If only a small number of delays are sufficient, they can supply useful information about the local history of the signal without raising much of a problem of shift invariance. Sensitivity to local time-distributed information can be encoded in the weights representing different delays.

The most important advantage of delay lines is that they allow inputs to be streamed. Unlike the previously described implementations of a time window, such a system can run continuously, with each input progressing down the row of delays and falling off the end. Such a design implies that learning a temporal template, that is, a pattern specified by absolute values of duration, should be easy to learn. For example, if voice-onset time (VOT) in syllables tends to be 50 ms, then the system should learn to look for a configuration of particular patterns at t_0 and at t_{-50ms}. Learning is adversely affected by variation in the time lag. That is, if voice onsets are distributed over the range from 25 ms to 100 ms, then the system will learn temporal blur across that range. Clearly, any normalization for tempo variation will have to be done at some other level. This design is ideal for recognizing absolute time values, hence they are a plausible wiring technique for extracting such information. For example, inter-aural time differences are important information for the angle of a sound source in the horizontal plane ([Shamma, 1989]).

Selected Feature Windows Another engineering model for temporal pattern recognition is based on portions of the frequency space. They track specialized kinds of events, by storing information about changes in the signal over time. Some models for perception of place of articulation in speech also exhibit this kind of basic structure ([Kewley-Port, 1983]). An example that is suggestive of a more general approach to temporal pattern perception is a model that detects slopes.

In this kind of model, activation of particular subnetworks indicates that specific spatio-temporal event sequence occurred. So a bank of detectors is constructed that span some range of frequency $\times$ time patterns. Time itself is no longer directly represented in the output of the detector circuit, since only some form of df/dt has been obtained. Depending on the temporal history of inputs and the current input, the system will adopt a particular unique state. In this way, the instantaneous state of the network can carry information both about what happened in the past as well as when it occurred. Recognition, prediction, decision-making and memory are not easily separable. The input arrives and, if it has the right properties, that is, if peaks occur in the right frequency regions, then activation is accumulated over time in a summing device. Errors in time of arrival cause no activations to be accumulated.

Study of neural structures in the visual system of a rabbit ([H. B. Barlow, 1965]) reveals frequency$\times$time detectors that measure the rate of change in various parts of the visual field. This idea has been explored in the connectionist paradigm by [Smythe, 1987,Smythe, 1988], [Watrous, 1990] and [Lang et al., 1990]. Smythe's system used a 'handwired' set of mini-networks for identifying the slopes of tracks in a binary-coded matrix. The model has a

battery of detectors for various positive and negative slopes across a range of 'frequencies'. In one order of firing of the input nodes all inputs would be summed. In a different order, a node could veto (or shunt) subsequent inputs. Competition between each slope detectors assures only one of these states being an attractor for a given input sequence.

This system illustrates one way that information distributed over time can be combined into a perceptual decision through use of delay nodes and detector systems based on shunting inhibition. The state of the system following presentation of a formant track carried information, not only about what the 'current input' is, but also about its history over the previous 5-10 slices. The system demonstrates the value of allowing very low-level decisions to be made that depend on the local time-history of the input. But these particular slope detectors, however, are rather inflexible. In a later section, a very different model that exhibits some of these properties will be described.

Hidden Markov Models. Hidden Markov models (HMMs) are the most effective speech recognition architecture currently available (for reviews see [Rabiner and Juang, 1986] [O'Shaughnessy, 1987]). In one application to an isolated-word speech recognition problem, there is a process whose 'hidden' states change about as often as phonemes do (that is, roughly 3 to 6 states for the optimum model of a consonant-vowel-consonant word). For each of these states, there is an 'observable' Markov process whose states are fixed-width spectra that have been clustered (or vector quantized) into a small codebook of spectral archetypes. Each of the two layers is Markovian since the probability of moving from state s_i to state s_j depends only on s_i. The matrix of probabilities for the spectrum layer also depends on the internal state. Thus, the model for each word (that is, the sequence of hidden states and transition probability vectors corresponding to it) can be used to find the probability that an observed sequence of spectra was generated by that model.

This kind of system is clearly dynamic in some sense, but some assumptions underlying these models limit their ability to accurately simulate dynamic cognitive processes. In general, the assumptions that make optimization possible limit the generality of the model. One example is the first-order Markovian assumption itself – that only one previous state influences the next state. Another is that the number of hidden states and layers can be fixed in advance. These topological constraints on the form of models that can be considered for stimulus patterns make them inappropriate as models for a general cognitive system. In addition, although the quantization process on the space of possible acoustic spectra seems essential in order to limit the size of the probability density vectors (given realistic constraints on available training data), this process also means that sensory discrimination is limited apriori. It seems that practical HMMs cannot have access to raw inputs. Hidden Markov models have many appealing properties including a tendency to automatically normalize for rate changes. But, of course, if the probability of word identity is compared across word models at conclusion of a test word, then a standard window is still assumed as a device for system control. Not all HMMs have this property (see [Levinson, 1985]).

The dynamic character of HMMs is a clue to the right direction to turn. Our final

class of systems does not save up the signal itself nor does it have a hardwired circuit or subnetwork that accumulates information about chosen inputs over pattern duration. Rather, it enters a trajectory when a known pattern begins to be presented. Events in the past change the state. But time is not totally squeezed out, since the representation still requires time to manifest itself.

2.2.1 Engineering Methods: General Observations.

Thus far, we have looked at systems that did one of two things. The linguistic models insisted that there is nothing of interest in time anyway. The engineering approach creates a window on the signal by saving it up, either externally to the system itself, or internally arrayed along delay lines. Then various programming and optimization techniques do whatever we may think of. This very practical approach is fine – until we want our data analysis technique to become a model of the perceptual process. At this point, the theoretical basis for data processing becomes a problem.

2.3 Dynamic Attractors

The third model for collecting information over time to support pattern recognition is a nonlinear dynamical system. Such a system can learn to be driven through a particular trajectory in its state-space by the input signals themselves. How to design such systems in connectionist networks is still not well explored – although developments are under way in various laboratories [Grossberg, 1982,Grossberg, 1986,Baird, 1986,Hirsch, 1989]. But there is some behavioral evidence (to be discussed below) that auditory memory for complex tone sequences take a similar form. If this is a general form for auditory memory, then such models might be useful for representing and perceiving human speech.

In addition, there is evidence that dynamic models are appropriate for modelling activity in the olfactory bulb. This work suggests that in an alert state, the activity of the olfactory bulb resembles a chaotic or random state [Skarda and Freeman, 1987,Baird, 1986]. When a very familiar odor is detected, the cells on the cortex of the bulb enter a state-space trajectory that is distinct for each odor. This limit cycle loops for several cycles until the stimulation provided by inspiration ends. These simulations have many similarities to those described below, althoug our modesl do not exhibit cyclic behavior.

2.3.1 Simple Recurrent Networks

Recently a number of investigators have attempted to deal with the weaknesses of feed-forward network models with an architecture that has a limited kind of recurrence – the simple recurrent network or SRN. The recurrence provides feedback and keeps track of history in some form. One example that runs continuously and stores stimulus history internally was explored by Elman.

Elman's *bi-daa-guuu* Network. An SRN was proposed [Elman, 1988] for tackling an interesting problem in sequence as shown in Figure 4. It is said to be an SRN because the activations of the hidden nodes, or sometimes outputs, are copied on the next cycle to the set of context nodes. Each node in the hidden layer sums inputs from outside and from the context nodes. Elman constructed many random strings of sequences *bi, daa* and *guuu* coded into 6-bit vectors (or 'distinctive features') for each character and presented one character at a time to the network. After optimization of weights on the feedforward edges, the system learns to predict most of the predictable properties of the strings. Its average error for the dimension that distinguishes the consonants from the vowels is low everywhere. For example, the system 'predicts' a consonant rather than a vowel after a single I, and also after the second A and after the third U. That is, the system has learned to predict structure in time based on the patterns that were presented. On the other hand, it predicts the place of articulation of a consonant very poorly at the boundaries between each substring, since these choices were made randomly.

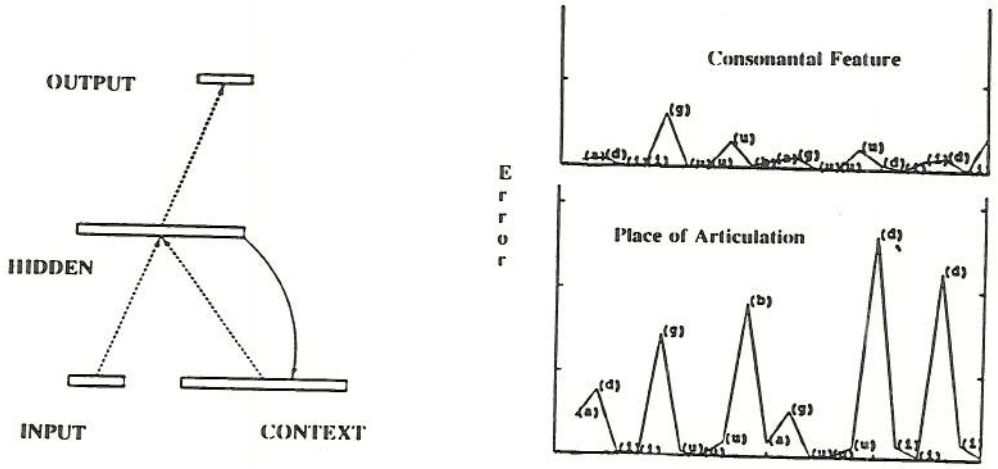

Figure 4: On the left, is Elman's recurrent network for the *bi-daa-guuu* problem. The input is fed one slice at a time. The hidden context node stores information about recent inputs. As shown on the right, the system can predict the output bit corresponding to the place of articulation feature of the consonants and vowels very well – except at the onset of each syllable where prediction cannot be done. The feature distinguishing consonants from vowels is always predictable.

Prediction is an essential task for networks that process variable length input patterns. This task can be acheived by employing a target-driven learning algorithm like back-propagation. The prediction target is what motivates the system to optimize toward representation of events in time. One important drawback of this system is that the inputs and the network itself run according to the same clock. Each input sample moves through the network in lockstep fashion. This clock reduces the ability of the network to recognize

patterns that vary in rate. Real nervous systems probably cannot depend on such a clock.

Recently, there have been numerous attempts to construct network systems that can learn to follow dynamic trajectories that differ depending on what input patterns arrive. These dynamic systems (a few examples are [Keeler, 1988,Mannes and Dorffner, 1989,Harris, 1989]) are intended to follow specified trajectories. To understand these models and their relation to the previous models, analytical techniques must be borrowed from the study of dynamical systems. If an energy field is parameterized by two dimensions of the model state space that constrains where the model will move in the next timestep, then we can conceptualize the instantaneous state of the system during a recognition problem as a ball rolling downhill in this 3-D 'landscape' [Hopfield, 1982]. The system is moving toward an attractive state. In the systems described above, the goal of design and training was to assure that the system would find the most attractive state (analogous to a least-energy state for a physical system). Hopefully, that state would represent the correct answer as an equilibrium point (or fixed point) of the system. On each trial (after training), the system 'falls' toward one of the target states. It can be guided in time by the input itself. These networks rely heavily on recurrent edges as the basic mechanism to store information and control the state. It also seems possible that higher order relations between nodes may be useful, that is, activation products can help assure rapid response. In our dynamic memory modules, there are fully recurrent cliques of nodes in which all nodes feed everything. In some systems, sigma-pi nodes allow the input and memory information to gate each other. In this way, successful prediction can be noted and made use of.

Recently Sven Anderson and I have been developing recurrent networks with dynamic attractors that learn to recognize melodies. For all the simulations reported here, two familiar tunes (shown in Figure 5) were recorded, 5 performances each, on an electronic keyboard (see [Port and Anderson, 1989]). The waveforms were spectrally analyzed in 64 ms Hamming windows with their centers 48 ms apart. Six frequency bins were centered around the 6 notes used in the pieces (between 261 Hz and 440 Hz, key of C). The average normalized amplitude in each bin was stored. Later, these files, one for each measure (with 8 spectra each), were fed, one slice at a time, to the 6 input nodes of various networks.

Tone Sequence Simulation with SRN. A recurrent network was designed, as shown in Figure 6 [Port and Anderson, 1989]. It has two outputs to train, a description of the next input frame (on the Prediction nodes) and a linearly ramped identification of the two targets. The hidden layer has sigma-pi nodes. By multiplying its 2 inputs (rather than just adding them), the context nodes can quickly amplify or attenuate their inputs (and vice versa).

We evaluated the ability of this kind of system to learn temporal patterns by training it to identify target measures from the music. Two measures, one from each tune, were selected as target melody fragments. The two target measures were particularly difficult to discriminate since they differed only in the duration of certain notes. Of the 5 recorded versions of each target measure, 2 were used as training tokens and 3 performances were reserved for testing. The competing 'distractor' measures were the 14 other musically

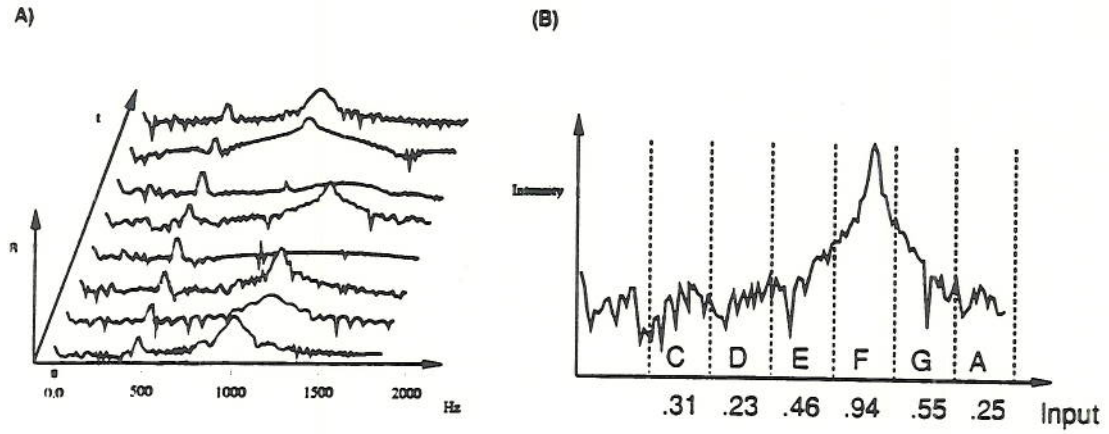

Figure 5: Stimulus construction for the melody recognition task. Panel A shows examples of the FFTs (fast Fourier spectra) for one measure. Panel B shows how a vector of 6 numbers was obtained by averaging across the frequency bin for each musical note.

distinct measures. The measures were presented in a continuous stream with no breaks between measures.

Networks of this architecture are not easily trained with backpropagation (since recurrent edges also were trained), so the "real-time recurrent learning" algorithm ([Williams and Zipser, 1
was used. This algorithm performs gradient descent in weight space but is computationally expensive (and, of course, not psychologically plausible). Despite variance due to the live performance, the system succeeded in recognizing the target measures among all the others, and in discriminating each target measure from all others with a d' of better than 2.5. [6]

Although successful, this network exhibits some limitations. When presented a simple task that required duration measurement, the system still made confusions between the two targets. This may be due to a weak representation of durational information. Further evidence of problems with the SRN design is that informal tests of changes in tempo (in which the rate of presentation of the target melodies was slowed down by presenting each frame twice), resulted in very poor performance. There is little reason to expect this model to handle rate change very well. In order to allow a dynamics that would be more flexible, the next network was designed. The idea was to build a fully recurrent clique of nodes and attempt to train the system to use dynamic changes in state to record the history of the

[6]The value d' measures discriminability unaffected by response bias. Values of d' are z scores along a hypothetical discriminant dimension along which *targets* lie toward one end and *distractors* toward the other end. Thus $d' = 3$ means the distributions lie of 3 standard deviations apart along the discriminant dimension. See [Swets, 1961,Robinson and Watson, 1972] or, for tutorial introduction, [Kantowicz and Sorkin, 1983].

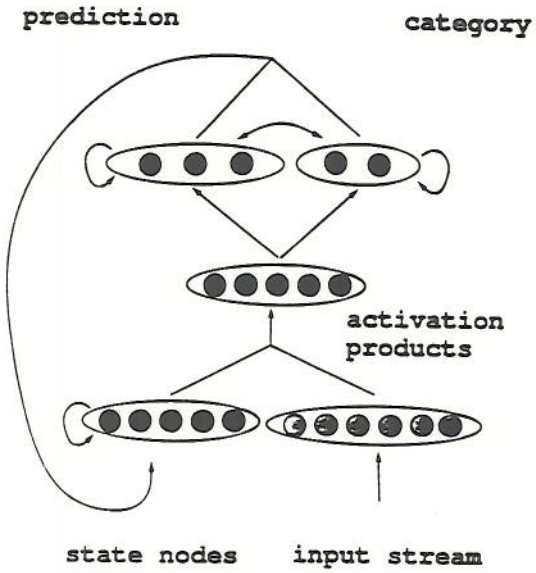

Figure 6: The SRN network in Port and Anderson, 1989. It was trained both to predict the next input on the prediction nodes and to identify the two target melody fragments on the category nodes by following a linear ramp target function. It succeeded in discriminating the two targets quite well.

input pattern.

2.3.2 Dynamic Memory Simulation

In our most recent simulations of the melody recognition, task, we have networks with a fully recurrent Dynamic Memory, as shown in Figure 7 (see [Anderson and Port, 1990] for more information). This system learned to recognize the melodies much better than the SRN model, getting d' of 3 (that is, 96% maximum correct identification assuming minimum response bias) for any pair of target measures.

More interesting than the basic result is the nature of the representation in the Dynamic Memory when a learned pattern occurred. Since there are 7 dimensions to be examined over time, it is convenient to perform *principle components analysis*. This rotates the 7 dimensions to find a small number of dimensions, the principle components, or PCs, that exhibit the most variance. Thus each PC basis vector is a linear combination of the 7 node activations. Thus Figure 8, shows the first two principle component values for continuous presentation of two target measures. For all the nontarget measures, the memory cycled around one region of its state space (that is, there was almost no differentiation between the distractors). These discrete trajectories provide estimates of the direction field in the basis provided by the principle components (similar to Waddington's 'adaptive landscapes' and 'energy landscapes'). This parameter represents a tendency to *move through the state space in a particular direction* (for an introduction to the ideas of dynamics, see

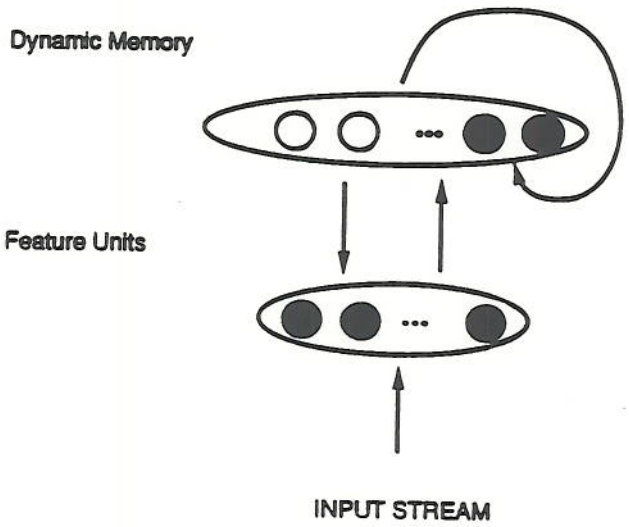

Figure 7: Dynamic network used for the melody recognition task in Anderson and Port, 1990. There are 5 input Feature nodes (the note A left out here to reduce the number of nodes) which feed to 7 fully connected Dynamic Memory nodes. Two of the memory nodes are trained to serve as identification nodes: one must turn on and the other off on the last 2 frames of the two targets. Real-time recurrent learning [Williams and Zipser, 1989] was used to optimize the weights.

[Abraham and Shaw, 1983] or [Luenberger, 1979]). When a target measure begins to be presented, the system moves through a very characteristic sequence of states as it 'falls' along a path in state space that is created by the effect of the stimulus input sequence itself and the dynamic memory. This system was gradually shaped by a slow learning process into a something that could respond dynamically so as to distinguish the patterns.

So 'recognition' of a trained pattern is exhibited in this short-term memory not by a single node lighting up (except for the two category nodes that were trained). Notice that the G in the target sequence GFED (Target 1) is far from the G in target sequence EEFG (Target 2). So the trajectory represents the whole pattern, not just their context-independent notes. That is, the trajectory and the instantaneous state at the end of the pattern both embody the whole pattern.

In order to see how well the system is following a trajectory, we should try to disturb it and see if it tends to cling to its trajectory. Although many possible ways to do this could be designed, one technique that we attempted was to slow down the rate of presentation. As shown in Figure 9, if the tempo of presentation is slowed down by a factor of 2 by simply presenting each spectral slice twice, the system still tracked the pattern and moved through the same regions of its state space – it just did so more slowly. The network was trained only on the standard tempo productions, but performed appropriately without retraining when the tempo was slowed by a factor of two. In other simulations it was shown that addition of noise on input amplitudes had little effect on performance unless very large.

Thus, the system is able to recognize patterns distributed in time even under certain

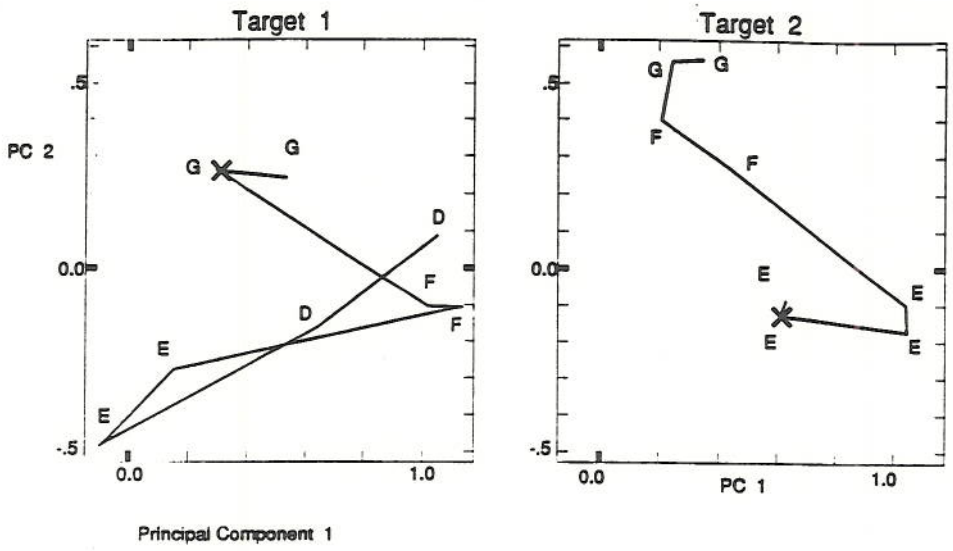

Figure 8: The first two principle components for the dynamic short-term memory node activations. The dark line follows the memory state during the 8 frames of pattern presentation. The X marks the second frame. For Target 1, the first two frames were a G, then two F's and so on. There were 8 frames for each target measure and they are connected in the order in which they were presented. Note that similar input notes (eg, G and F have very different locations in each pattern. All the nontarget measures lie in a single small attractor near the center of the space. See Anderson and Port, 1990 for more information).

distortions of time and frequency. It was able to differentiate patterns by recursively exciting itself such that, as long as inputs continue to support the pattern, then it follows a trajectory through its state space. This trajectory was followed despite major distortions of duration. Presumably, it will also exhibit resistance to other kinds of noise. Further exploration of the resilience of this kind of system is under way.

This kind of representation appears to have certain advantages for a nervous system.

1. It has a built-in mechanism for reset when a pattern is completed. That is, it should be able to learn patterns of any duration. The outputs shown above plus results of other unpublished simulations suggest that fully connected network cliques trained with real-time recurrent learning can allow inputs to control their own processing in a reasonable and stable way.

2. Response can be initiated as soon as possible, that is, as soon as the information in the stimulus permits a decision to be made.

3. Tempo invariance of the system was obtained 'naturally', that is, without training on more than a single tempo. The tendency to follow a trajectory provided a considerable amount of tempo invariance. Presumably there are limits on human tolerance of complexity in tempo change which should be investigated experimentally.

4. Finally, the system reaches recognition without having to label or identify any components. It gets a label for the entire sequence but the label depended on generation of a memory that was temporally stable. It employed a 'subsymbolic' representation of temporal pattern to enable symbol-like recognition of the pattern as a whole.

The model has other implications, however. Scaleup of this idea seems to imply learning a large number of state trajectories for complex auditory patterns of various sizes and complexity. In addition, these learned trajectories will have to nest within each other hierarchically. The important implication of this model for a psychological theory of time is that there is nothing that directly represents time. Thus there is no 'chronotopic field' such that distances along it represent intervals of real time. Instead, the lowest level of the system does 'time normalization' intrinsically and automatically. It directly extracts predictable patterns from inputs, and stores them in a useful way. But to get such representation requires much practice in order to tune the dynamics of memory. A learning process is clearly required to support development of a model for the states themselves. Actually, as will be argued in the next section, there is evidence that human perception depends on similar kinds of learning for the lowest level of auditory pattern perception.

3 Behavioral Evidence: Discrimination of Complex Sound Patterns

One conclusion from this review of temporal pattern recognition techniques is that static, timeless models, of the kind used in linguistics, are quite inadequate for perceptual models. And time windows, which are implicit in all research on speech and hearing, are implausible biologically. Many kinds of experiments have been interpreted as support for a short-term echo-like store. In these experiments subjects typically listen to stimuli that are either very simple or else very familiar. The important question, however, is whether there is a raw *acoustic* store in which some parameter represents time itself. Is there a true *chronotopic field* containing spectra displayed through time which subjects can examine? To explore this issue, subjects must be challenged with complex and unfamiliar patterns. Charles Watson and his colleagues have conducted a large number experiments on such patterns over the years (for reviews see [Watson and Foyle, 1985,Watson, 1987,Espinoza-Varas and Watson, 1986]). Subjects' performance on these tasks does not seem to be encouraging for a model that assumes access to a spectrogram-like store of the raw acoustic signal.

The basic idea of this research program was to make very complex acoustic patterns and evaluate subjects' ability to make detailed judgments about the patterns. That is, in the terminology of cognitive science, they explored the *quality of auditory representation.* Subjects were presented two very similar sequences and asked if they are the same or different, as shown in Figure 10. The quality of the representation is measured as the proportional change in the stimulus that is required for subjects to reach a threshold percent correct (usually 70%) discrimination of the difference. A typical stimulus in these experiments is a random sequence of 10 tones, each of which is only 40 ms long. Thus, a stimulus is a complex burst of sound (vaguely resembling a turkey gobble) that lasts less than a half a second. The subject is played one of these sequences as a standard and then

twice more, and asked if the last two are identical. One of the last two items may have one component changed somewhat in frequency or, in some experiments, duration or intensity. The primary dependent variable, then, is the $\Delta s/s$, that is, a percentage change in a stimulus parameter, that is required in order to reach a certain level of performance. The threshold Weber fractions for the parameters of frequency, intensity and duration are well known, of course, for very simple tone stimuli. The question raised in these experiments is what limitations on performance lie central with respect to the ears? If subjects need to store a complex pattern, what is the representation like? How much detail do listeners have access to?

The general finding is that if the random patterns presented to a subject are drawn from a large set (eg, more than 50 patterns), then subjects are unable to do this task very well. Even after thousands of trials, stimuli that are complex and unfamiliar require enormous changes in stimulus parameters for the differences to be detectible [Watson and Foyle, 1985]. For example, in many cases frequency must be changed by more than an octave (that is, by a factor of 2) before subjects can detect the change reliably. These changes are large enough that a temporally integrated representation, that is, something like a long-time spectrum averaged across the entire sequence, should be able to reveal the difference.

On the other hand, if the subject is trained on one pattern at a time, then his performance at detection of a frequency change gradually approaches his performance on the tones if they were presented under ideal experimental circumstances (where the tones are long and immediately adjacent in time). However, asymptotic performance still may take several thousand trials on each pattern [Spiegel and Watson, 1981,Leek and Watson, 1984]. It is possible that the training involved here produces a similar representation to that achieved by the Anderson and Port network described above. That is, training may produce distinct dynamic trajectories through state space for learned patterns.

There are several kinds of evidence that support this account. The first is that it appears that when subjects listen to patterns, correction for tempo variation is obtained quite naturally, almost for free [Kidd and Watson, 1989]. It may initially seem counterintuitive, but slowing down these very complex patterns by a factor of 4 or 8 provides almost no improvement in performance [Watson and Foyle, 1985]. Even abrupt and unpredictable changes in tempo do not impair performance whereas changes in the frequency range of a pattern impairs performance greatly ([Kidd and Watson, 1989]). These characteristics are compatible with the view that for rapid complex patterns, the auditory system learns dynamic trajectories through state space. As the pattern appears at the periphery, the system is driven through this trajectory. Some deviations in the pattern, such as changes in the overall rate of presentation should not disturb it very much. Thus, these behavioral results resemble our dynamic memory system. Both require extensive training but are resistant to rate changes and should not exhibit backward masking (since patterns are recognized *as they come in*, not after a period of subsequent processing).

In one experiment, Watson's subjects were given many patterns – but with a special constraint in one of the conditions: that Δf was always added to the 7th component of each 10-tone pattern. The subjects were not able to exploit this invariant to perform better

here than they did in the condition where the target tone (the one with Δf) moved around from pattern to pattern [Watson et al., 1975]. So they were apparently unable to ignore all other tones in the pattern except the 7th. They had to try to learn all the rest of the pattern too. These results suggest that a *temporal mask*, one specified in absolute values, is more difficult to learn than a 'mask' that is based on ordinal position. Conversely, patterns that are defined in tempo-invariant terms (like a melody or word) should be difficult for the delay-line model. A dynamic model should have a difficult time recognizing events at a fixed time lag when intermediate events are varied. The intrinsic time normalization and trajectory tracking of this system should reduce performance on such tasks. Obviously, these are empirical issues that should be explored more carefully experimentally.

4 Conclusions

It is proposed that auditory memory suitable for recognizing long time-window patterns may be achieved by continuously recoding the signal into descriptions that summarize useful properties of what has been seen recently – including temporal properties. That recoding process is achieved by making decisions (either continuously or on each input time frame) as to what description is appropriate. Processing is controlled primarily by the input sequence itself in conjunction with long-term learning (stored in the weights) and dynamic responses to recently seen inputs. This means that control over the analysis process is inherently bound up with the description itself, and that data structures are not clearly distinguishable from the operations upon them. If recodings of this kind at various time scales and various levels of abstractness can be constructed, then perhaps continuous recognition of hierarchically-structured dynamic signals like speech is possible. But success may require abandoning the common assumption of cognitive science that everything in knowledge should have a representation. A clear distinction between operations and representations may be impossible to make in a dynamic model. Instead of concatenation, there will be a trajectory through state space. Instead of static symbols, there will be only the stability of equilibria.

The words of human speech always exist in real time – in historical time. But the formal symbolic model that offers so much power and clarity to modern thought about cognition does so by giving historical time short shrift. Approaching the problem from a formalist perspective, the assumption once seemed almost inescapable that scalar properties of time can either be ignored and processed as serial order, or else, if that fails, that time can be treated simply as another parameter. But if it is a parameter, then it must be measured. And the measurement task places a huge burden on some preprocessing system that must both display time as space (or is there some *other* way?) and recognize certain landmarks and measure the required values as rational numbers. If our concern is with the description of sound, then we should go ahead and measure time. But if measurements are supposed to play a role in a model of cognition, then we are responsible to account for the extraction of this information. There are only a few options. What I have tried to show in this essay is the narrow range of underlying models that have actually been exploited in the many

disciplines with a concern for speech. I have also tried to suggest that another model is possible, one that is intrinsically dynamic.

Although the implementations in this paper are novel, the basic ideas have been on the table for many years – especially in neuroscience. For example, Lashley's informal model of cortical processing of patterns in time ([Lashley, 1951]) sounds fresh and relevant today:

> "The cortex must be regarded as a great network of reverberatory circuits, constantly active. A new stimulus, reaching such a system, does not excite an isolated reflex path but must produce widespread changes in the pattern of excitation throughout a whole system of already interacting neurons."

To understand things, we prefer to make them stand still for us. But cognition may be something that can only be understood by plunging into models that are based on 'reverberation'.

References

Abraham, R. and Shaw, C. (1983). *Dynamics, The Geometry of Behavior, Part 1.* Aerial Press, Santa Cruz, California.

Anderson, S. and Port, R. (1990). Network model of auditory pattern recognition. Technical Report 11, Indiana University, Cognitive Science Program.

Baird, B. (1986). Nonlinear dynamics of pattern formation and pattern recognition in the rabbit olfactory bulb. *Physica*, 22D:150–175.

Bever, T. G. (1973). Serial position and response biases do not account for the effect of syntactic structure on the location of brief noises during sentences. *Journal of Psycholinguistic Research*, 2(3):287–288.

Bregman, A. S. and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89:244–249.

Carlson, R. and Granstrom, B., editors (1982). *Representation of Speech in the Peripheral Auditory System.* Elsevier; Amsterdam.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* Harper and Row, New York.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, 2:223–274.

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

Dorman, M., Raphael, L., and Liberman, A. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, 65:1518–32.

Elman, J. (1988). Finding structure in time. Technical Report 8801, Center for Research in Language, University of California at San Diego, La Jolla, CA.

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

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

Espinoza-Varas, B. and Watson, C. (1986). Temporal discrimination for single components of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, 80(6):1685–1694.

Fant, G. (1973). *Speech Sounds and Features*. MIT Press, Cambridge, MA.

Gasser, M. and Lee, C.-D. (1989). Networks that learn phonology. Technical Report 300, Computer Science Department, Indiana University.

Goldsmith, J. (1976). *Autosegmental Phonology*. Garland Press, New York, NY.

Grossberg, S. (1982). *Studies of Mind and Brain*, volume 70 of *Boston Studies in the Philosophy of Science*. D. Reidel Publishing Company, Dordrecht, Holland.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech language, and motor control. In Schwab, E. and Nusbaum, H., editors, *Pattern Recognition by Humans and Machines: Speech Perception*. Academic Press, Orlando, Florida.

H. B. Barlow, W. R. L. (1965). The mechanism of directionally selective units in a rabbit's retina. *Journal of Physiology*, 173:477–504.

Halle, M. and Stevens, K. N. (1980). A note on laryngeal features. *Quarterly Progress Report, Research Lab of Electronics, MIT*, 101:198–213.

Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Bradford Books/MIT Press, Cambridge, Mass.

Hare, M. L. (1990). The role of similarity in hungarian vowel harmony: A connectionist account. *Connection Science*.

Harris, C. L. (1989). Connectionist explorations in cognitive linguistics. Technical report, UCSD.

Hinton, G. (1988). Representing part-whole hierarchies in connectionist networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 48–54, New Jersey. Erlbaum.

Hirsch, M. W. (1989). Convergent activation dynamics in continuous time network. *Neural Networks*, 2:331–349.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558. National Academy of Sciences.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23:67–72.

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.* MIT Press, Cambridge, Massachusetts.

Kantowicz, B. and Sorkin, R. (1983). *Human Factors: Understanding people-system relationships.* Wiley, New York.

Keeler, J. (1988). Comparison between Kanerva's SDM and Hopfield-type neural networks. *Cognitive Science*, 12:299–329.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 73:322–335.

Kidd, G. R. and Watson, C. S. (1989). Detection of changes in frequency- and time-transposed auditory patterns. Paper presented at the Psychonomics Society.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59:1208–21.

Klatt, D. (1986). Problem of variability in speech recognition and in models of speech perception. In Perkell, J. and Klatt, D., editors, *Invariance and Variability in the Speech Processes*, pages 300–320. Erlbaum Associates, Hillsdale, NJ.

Ladefoged, P. (1989). Representing phonetic structure. Working Papers in Phonetics 73, University of California, Los Angeles.

Lakoff, G. (1988). Cognitive phonology. Paper presented at the LSA Annual Meeting.

Lang, K. J., Waibel, A. H., and Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23–43.

Lashley, K. S. (1951). The problem of serial order in behavior. In Jefress, L. A., editor, *Cerebral Mechanisms in Behavior*, pages 112–136. Wiley, New York.

Lea, W. A. (1980). *Trends in Speech Recognition.* Prentice Hall, Englewood Cliffs.

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

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

Levinson, S. E. (1985). A unified theory of composite pattern analysis for automatic speech recognition. In Fallside, F. and Woods, W. A., editors, *Computer Speech Processiong*, chapter 9, pages 243–272. Prentice-Hall International.

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

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.

Luenberger, D. G. (1979). *Introduction to Dynamic Systems.* Wiley, New York.

Mannes, C. and Dorffner, G. (1989). Self-organizing detectors for spatiotemporal patterns. Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, Vienna, Austria.

Moore, B. C. J. (1982). *An Introduction to Psychology of Hearing.* Harcourt Brace Jovanovich, second edition edition.

Neisser, U. (1967). *Cognitive Psychology.* Appleton.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine.* Addison-Wesley, New York.

Port, R. (1986). Invariance in phonetics. In Perkell, J. and Klatt, D., editors, *Invariance and Variability in Speech Processes*, pages 540–558. Erlbaum Associates, Hillsdale, New Jersey.

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

Port, R. and Crawford, P. (1989). Pragmatic effects on neutralization rules. *Journal of Phonetics*, 17(4). To appear.

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

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

Port, R. F. and Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America*, 66(3):654–662.

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–16.

Repp, B. (1984). Categorical perception: issues, methods and findings. In Lass, N. J., editor, *Speech and Language: Advances in Basic Research and Practice, Vol. 10*, pages 243–335. L. Erlbaum.

Robert Port, W. R. and Maki, D. (1988). Use of syllable-scale timing to discriminate words. *Journal of the Acoustical Society of America*, 83(1):265–273.

Robinson, D. E. and Watson, C. S. (1972). Psychophysical methods in modern psychoacoustics. In Tobias, J. V., editor, *Foundations of Modern Auditory Theory*, volume 2, chapter 3, pages 99–131. Academic Press, New York, New York 10003.

Sachs, M. B. and Young, E. D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *Journal of the Acoustical Society of America*, 68:858–875.

Sankoff, D. and Kruskal, J. B., editors (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, Mass.

Sejnowski, T. and Rosenberg, C. (1986). Nettalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, The Johns Hopkins University Electrical Engineering and Computer Science.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanisation of thought processes*, pages 511–531. London: H. M. Stationery Office.

Shamma, S. A. (1989). Stereausis: Binaural processing without neural delays. *Journal of the Acoustical Society of America*, 86(3):989–1006.

Skarda, C. and Freeman, W. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, 10:161–195.

Smythe, E. (1987). The detection of formant transitions in a connectionist network. In *Proceedings of the First IEEE International Conference on Neural Networks*, pages 495–503, San Diego, California.

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

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(1):223–230.

Stevens, K. N. (1983). Design features of speech sound systems. In MacNeilage, P., editor, *The Production of Speech*, pages 247–262. Springer-Verlag.

Stevens, K. N. and Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In Eimas, P. and Miller, J., editors, *Perspectives on the Study of Speech.* L. Erlbaum, Hillsdale, NJ.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S., editor, *Handbook of Experimental Psychology*, pages 1–49. Wiley, New York.

Swets, J. A. (1961). Is there a sensory threshold? *Science*, 34:168–177.

Tank, D. and Hopfield, J. (1987). Neural computation by concentrating information in time. In *Proceedings of the National Academy of Sciences*, pages 1896–1900.

Vassiere, J. (1985). Speech recognition: A tutorial. In Fallside, F. and Woods, W. A., editors, *Computer Speech Processing*, chapter 8, pages 191–242. Prentice Hall International.

Waibel, A. (1986). *Prosody and Speech Recognition.* PhD thesis, Carnegie-Mellon University, Computer Science Dept. Pittsburgh, Pennsylvania 15213.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1988). Phoneme recognition: Neural networks vs. hidden Markov models. In *Proceedings of the ICASSP*, pages 107–110. IEEE.

Warren, R. and Bashford, J. (1981). Perception of acoustic iterance: pitch and infrapitch. *Perception and Psychophysics*, 29(4):395–402.

Watrous, R. (1990). Phoneme discrimination using connectionist networks. *Journal of the Acoustical Society of America*, 87:in press.

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.

Watson, C. S. (1987). Uncertainty, informational masking, and the capacity of immediate auditory memory. In Yost, W. A., editor, *Auditory Processing of Complex Sounds*, pages 267–277. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Watson, C. S., Wroton, H. W., Kelly, W. J., and Benbasset, C. A. (1975). Factors in the discrimination of tonal patterns. I. component frequency, temporal position, and silent intervals. *Journal of the Acoustical Society of America*, 57:1175–1181.

Wheeler, D. and Touretzky, D. (1989). A connectionist implementation of cognitive phonology. Technical Report CMU-CS-89-144, School of Computer Science, CMU.

Williams, R. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

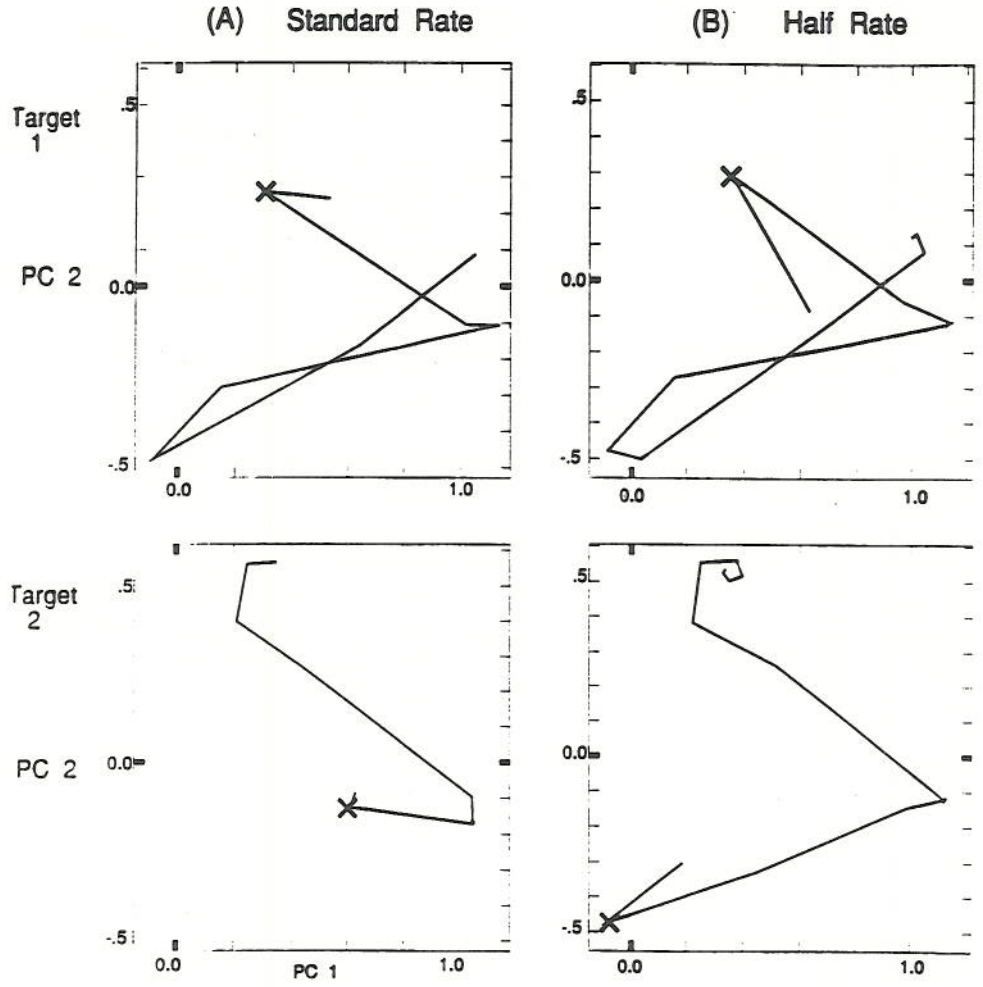

Figure 9: Effects of changing tempo. On the left is the same display as the previous figure. On the right the same pair of stimuli were presented to the network at half-tempo by repeating each spectral frame twice. Almost same trajectory was followed for the slow presentation as before, but it now takes 16 frames to complete the trajectory instead of 8. Notice for Target 2 in the Slow presentation that the system circles closer to the final attractor state. No retraining was done for this response.

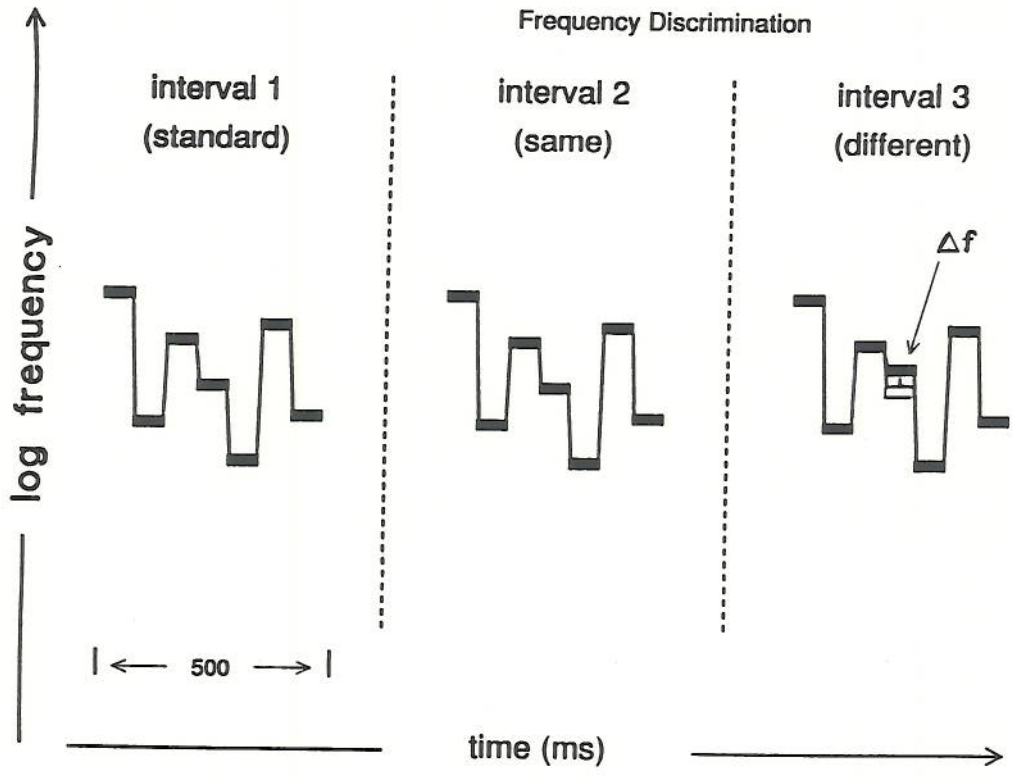

Figure 10: Sample trial for discrimination of a frequency change in a three-interval task. The first is the standard, then either the second or third item may have Δf added to one component. Unless the subject is very familiar with the pattern this task is very difficult. If the patterns are unfamiliar, critical $\Delta f/f$ (the percentage change to permit good discrimination) must be an order of magnitude larger than if the pattern is highly trained.