TECHNICAL REPORT NO. 265

Representation and Recognition of Temporal Patterns

by

Robert Port

January 1990

COMPUTER SCIENCE DEPARTMENT
INDIANA UNIVERSITY
Bloomington, Indiana 47405-4101

Representation and Recognition of Temporal Patterns

Robert Port[1]
Department of Linguistics,
Department of Computer Science
Indiana University, Bloomington, Indiana 47405

January 15, 1990

[1]I am grateful to Sven Anderson for major contributions to the basic ideas described here as well as to the research. I am also grateful to Michael Gasser, Charles Watson, Gary Kidd, Jungyul Suh and John R. Merrill for helpful discussion of these ideas. This research was supported in part by the National Science Foundation, Grants DCR-8505635 and DCR-8518725, and by the Air Force Office of Scientific Research.

Abstract

How can a nervous system represent for itself the temporal relations of patterns that it knows? For auditory patterns, the nervous system must obtain a pattern that is distinctive and reliably reproduceable. This paper reviews 4 basic models that span the range of theories addressing this problem. Many of these have been implemented in connectionist networks. In linguistic models, for example, (1) there are only *ordered symbols*, like phonemic segments, that occur in a particular order, where the only measure of length is the number of symbols. In traditional speech recognition techniques (2) a *static time window* that stores the entire signal and allows measurement of whatever intervals one might think of. There is very little evidence supporting a time window of relatively unprocessed acoustic information. There are also(3) *selected feature windows* that look for changes in input within a particular portion (either hardwired or learned) of the input spectrum. Finally, there is (4) a *dynamic system* as a connectionist network. Such models are driven through a particular trajectory in state-space by the input signals themselves. The representation does not distinguish frequency from time very well. Design of such systems is still very problematic, but there is some evidence that auditory memory for complex tone sequences may be of this form. If this is the general form for auditory memory of dynamic patterns, then such models might be useful for representing and perceiving human speech.

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 dance rhythm you are familiar with? Apparently, the central nervous system reaches some stable state that is distinctive to the pattern itself. To reach this state of recognition, requires a representation for the stimulus over time that is sufficient to permit recognition to occur. So there are two basic issues. How is early information recorded to permit recognition of something distributed in time? And secondly, what is the structure that is 'active' when a pattern is recognized? The general issues surrounding these questions about the representation of time have not attracted a great deal of attention in the literature. Elman [Elman, 1988] has addressed the question quite directly and his simulations have explored a variety of models that will be mentioned below. Others in many disciplines have also addressed the question in various ways. This paper reviews 4 important models for the representation of sound as events in time. They are not claimed to comprise a typology of models, but hopefully they span the range of well-known theories addressing this problem. After this attempted typology, the paper will discuss evidence from research on the perception of complex tone patterns. These data suggest that one of the models, the dynamic model, is supported by actual limitations on the ability of humans to learn very complex tone 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 models addressing problems with 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 in cognition.

The first model for temporal representation to be discussed below is the standard view of the symbolic structure of cognition. This model is nicely illustrated in linguistic theories about words. In classical linguistics, words were said to be spelled as ordered strings of segmental phonemes – consonants and vowels. Recent thinking in phonology [Goldsmith, 1976,Clements, 1985] has expanded this model to include 'autosegmental' data structures that allow each articulatory system to define its own time scale. But the autosegmental tiers are still linked together in a single spatio-temporal structure. So this model at least treats time no longer as an integer scale [Stevens, 1951], but it is still ordinal. That is, the only notion of time is serial order.

Two analogs of realworld temporal order can be defined within the symbolic model: first, locations in a buffer precede and follow one another, and secondly, states of a derivation precede and follow one another in 'derivation time' – that is, in the non-historical, timeless time in which mathematical operations are carried out. [1] The approach endorsed in this paper, however, is to abandon the traditional question: "How is it that temporal patterns can be viewed as special cases of ordered symbols?" Instead it should be replaced with the question "How is it that symbol order patterns can be viewed as special cases of more general patterns in real time?"

The second important model for temporal pattern recognition is the time window model in which any kind of measurement may be made on the signal distributed in space across a buffer. In

[1] The treatment of derivation states as time sounds absurd in artificial intelligence. Noone would 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 approaches to phonology inspired by connectionism have attempted to have rules apply simultaneously ([Lakoff, 1988] or simultaneously except in special cases that make use of special purpose hardware [Wheeler and Touretzky, 1989].

this framework, much more information about events in time can be extracted and represented as numerical parameters. Quantitative relations can then be defined such as 'the relative duration of X and Y'. But this power is achieved at the cost of assuming the possibility of a time buffer, plus preprocessing to locate points in time for labelling.

The third technique selects certain frequency regions and reports on particular kinds of changes in signal over that region. The fourth technique and the most subtle to design is a full dynamical model that learns to respond dynamically to a particular dynamic pattern. This type may reach a unique state for each trained pattern, but it does so as part of a trajectory. The representation of time here seems to be both a trajectory, that is, a dynamic pattern, and a state that represents the dynamics of the signal itself.

These four models or frameworks each 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.

2 The Problem of Temporal Patterns

What is meant by a *temporal pattern* for the study of cognition? All sound consists of variations in air pressure over time, yet some patterns can be quite long even though we may 'know' every detail. Should we require temporal patterns to have a particular kind of internal structure (like periodic rhythms)? I don't think so. For our purposes, both a word and a sentence as uttered are temporal patterns. For sentences, it is true that the most important structures can be expressed as ordered symbols (words or morphemes) for which temporal detail is not a very relevant property. But ordered morphemes are really only a special case of a temporal structure that happens to allow a wide range of invariance transformations ([Port, 1981,Port, 1986]). In conventional linguistic theory, time constitutes an *ordinal scale*, in the familiar terms of S.S. Stevens [Stevens, 1951], since serial order is the only property that remains invariant under permitted transformations. This seems intuitively correct for abstract linguistic levels (like syntax), but at shorter time scales (like for quasi-rhythmic units such as syllables and segments), the assumption that serial order is all that is cognitively important seems misguided. Many temporal patterns, of course, have periodic structures at multiple hierarchical levels – as in musical rhythms. But, for the purposes of cognitive science, periodic and quasi-periodic patterns should be viewed as equally temporal. Both recur in time with a structure that is only partially predictable.

So, the temporal patterns that are the concern of this paper may be periodic (like music), quasi-periodic (like speech and language) or isolated non-repeating acoustic events with complex internal structure (like chirps or thumps). It does not seem that there is a theoretical limit on the maximum duration of a temporal pattern.

The most difficult part of the problem of defining the concept of a temporal pattern is the question of how short they can be. The highest audible tone (20,000 Hz) has a period of one twentieth of a millisecond. Are all sounds to be viewed as temporal patterns, then? No. Since our interest is in central processing of sound, we need to model the acoustic signal as it is represented in the auditory nerve. In the remainder of this section, some of the basic facts of hearing are reviewed which argue for a particular lower bound on the duration of temporal patterns (see [Handel, 1989] for an accessible survey of audition.)

A lower bound on the duration of temporal patterns can be roughly defined from study of the form of auditory information in the auditory nerve. The central nervous system receives continuous information about acoustic events in the form of the firing rate of a large number of neurons. High frequency tones, such as a 10,000 Hz tone, mechanically excite the basilar membrane so that after several cycles of input, the waves travelling down the basilar membrane reach an amplitude maximum at a region of the membrane that is specific to 10,000 Hz. Due to frequency sharpening by lateral inhibition, only cells in the auditory nerve with center tuning-frequencies at the disturbance maximum transmit a signal. Thus for frequencies from 20,000 Hz down as low as 200 Hz there are particular cells in the auditory nerve tuned to their frequency. Integrated rate of firing in these cells is proportional to the intensity of sound near their center frequency. This means that frequencies above 200 Hz are presented as independent and parallel channels that present continuous information to the central nervous system. These channels can be modelled as amplitude values for a small number of functionally independent channels. There are about 30 critical-band channels in the audio range, each about a 1/3 octave wide. In very simple terms then, over this range sound is presented to the central nervous system in a form somewhat like a continuously streaming sound spectrogram.

On the other hand, low frequency sounds, those below 200 Hz, do not produce an amplitude maximum anywhere on the basilar membrane. They are represented only as temporal events. Any pattern that repeats less often than about 200 times a second must therefore be presented to the central nervous system as a signal in time. Frequencies between about 200 Hz and 1000 Hz have both kinds of representation - as independent frequency channels and as activation bursts synchronized to a particular phase angle of the incoming signal. Across the range of overlap there is a graded increase and decrease of dependence on the two modes of representation.

The simplest model of the auditory signal suitable for the study of auditory temporal patterns would be to employ spectral samples made at twice the rate of 200 Hz (since only a 400 Hz sampling rate could reproduce a 200 Hz pattern without aliasing). This implies spectral sections with a width of 2.5 ms. [2] For the moment, we can model the auditory signal with brief fast-Fourier spectra (FFT spectra) sampled with non-overlapping sections of 5 ms. This is, in fact, close to the representations used by many contemporary models for speech recognition ([Klatt, 1986]) and it has many similarities to the display of a standard wide-band sound spectrogram (which has a 3 ms integration window). In short, we can say that a temporal pattern for cognitive science must be representable with something between 100 and a 1000 spectral frames per second.

The scope of this essay is to review several well-known models for the representation of temporal patterns within a system designed for auditory pattern recognition – whether those models were designed as putative models for cognition or intended merely as engineered devices. I will consider each critically for its practical strengths and limitations keeping in the foreground the requirements of animal nervous systems.

3 Temporal Representations

In the following four sections, various models for representing temporal patterns in a perceptual or recognition system are sketched. These models have been popular within different disciplines

[2] Of course, it is possible that for the dynamics of a central model to work correctly, a satisfactory discrete model for continuous auditory inputs would require a higher effective sampling rate than this minimum value.

dealing with speech since they seemed to address the necessary issues. The first technique manifests the underlying view that serial order is where essential linguistic information about speech sounds can be found. The second one allows the perceptual system to simultaneously look at the present and some amount of the past and permits unlimited use of information in the window. The third technique postulates frequency or other selectivity on inputs, while the fourth exhibits dynamic trajectories in state space to distinguish different temporal patterns.

What representation of the signal should there be in order to support perception of temporal patterns? Very long and abstract patterns in this space – such as a familar melody – should probably be represented in abstract units like the notes of a symbolic musical scale. The most problematic issues arise for short duration patterns, like stop consonants and syllables – not to mention chair squeeks, light switch clicks and refrigerator door thumps, cat meeows, etc. These patterns are short enough that there may be no 'symbol-sized pieces' from which they could be plausibly represented. It is this initial code for temporal patterns that must be modeled first before we will succeed in finding the higher-level units like notes of a musical scale and words. And besides, it is unlikely that a representation of a melody or a word can really ignore *everything* about the temporal pattern. For example, changes in temporal detail can change one word into another (eg, [Port and Crawford, 1989]. Thus, the temporal invariance of words *as pronounced* is clearly not so abstract and unconstrained in permissible timing distortions as is implied by the notion of an ordinal scale implies (see [Dorman et al., 1979,Port, 1986,Port and Dalby, 1982,Port and Crawford, 1989]. These are some of the empirical issues surrounding the question of what kind of representation of events in time is required to model cognition. In any case, this tour begins at the simplest symbolic theory – with linguistics. For linguistics time is clearly only an ordinal integer scale.

3.1 Ordinal Symbols

In a strictly symbolic system like linguistics, there can be nothing but ordered symbols. In linguistic models, that is, in phonology, the sequentiality of words can be expressed by 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]). The only possible measure of distance in time within such a system is in terms of the number of symbols (though, of course, symbols can be defined that are intended to mean 'longer' and 'shorter' (See [Lisker and Abramson, 1971]). But acoustic inputs are actually distributed in real time, so the 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. 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 difficult to imagine how any perceptual system could work if it did not eventually integrate events that are close enough together.) This integration process necessarily results in loss of information about the precise distribution 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.

3.1.1 Jakobson's Distinctive Features

Since Jakobson's early work, feature-detector systems have normally 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. 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. 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 from real-time integration of the auditory signal to a serially ordered symbolic structure for language and other higher cognition.

This perceptual model was conceptualized by [Jakobson et al., 1952] 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 within the signal buffer that they implicitly assumed.

Most of the 12 features in the system of Jakobson, Fant and Halle are defined in terms of information integrated over specific time windows of 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 feature), the stop would be categorized as plus or minus for each feature. Thus, the linguistic features from which phonemes were defined were 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 was never actually implemented in a speech recognition device, the conceptual model thrived 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 spectral templates [Stevens and Blumstein, 1981,Stevens, 1983]. Other models, however, (eg, [Kewley-Port, 1983]) have emphasized that performance can be improved if a role is allowed for dynamic properties of the signal. 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 'more temporal' than simply timeless, ordered features needs to be considered.[3]

Engineers, trying to do speech recognition found temporal information to be primarily troublesome rather than helpful. So they chose to attempt a mapping from a spectrally sampled speech

[3] Another prominent example of such a dispute between phoneticians and linguists about the role of speech timing 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 is an important cue for speech perception. 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 temporal information cannot be ignored [Lisker and Abramson, 1971]. The issue was never resolved, but apparently simply dropped.

signal onto sequential representations in which serial order is the only invariant.

3.1.2 The Time Warp Model

Over the past 15 years or so, researchers with engineering interests have developed a more practical means of modelling speech for the specific purpose of recognizing isolated words. Despite the evidence of the role of temporal patterns in the specification of words, it appeared that this information could not be incorporated in speech recognition very easily [Waibel et al., 1988]. So speech recognition techniques since the mid-1970s (as reviewed in [Lea, 1980] and [Vassiere, 1985]) make a buffer of signal samples (represented as spectra of some sort or other). Various filters (eg, zero-crossings in the waveform, energy in various frequency bands, etc) are convolved with the buffer to generate a vector for each time slice in the utterance. Pattern matching is done by first comparing the vector of the test item with each vector in the stored template. Since timing will typically vary between template and test item, these temporal discrepancies contribute greatly to the total error for the comparison. Thus the correct word often gets a very high error score when the timing is different from the template. The solution employed is dynamic time-warping [Itakura, 1975,Sankoff and Kruskal, 1983]. This technique, in its simplest form, allows comparisons to be based almost entirely on serial order (though many variants of the technique have been tried).[4]

Time warping systems construct a buffer of the waveform. Primitive spectra and various other measures were computed every 5-20 ms. After collecting and averaging a number of 'training tokens' of each word in the vocabulary, the system stores a 'template' for each word against which new test tokens can be compared. Up to this point, time is treated as an integral number (depending on the number of spectral samples). However, since test items can occur at a wide variety of rates and styles, time must be normalized. But a technique that simply scales all durations by a constant amount does not work well. So, dynamic time warping, the strongest possible nonlinear transformation of time, as shown in Figure 1, is done by finding a minimum error mapping from the template vector to the test item vector. In its simplest form the vector for each time slice is subtracted from all the vectors in the template. The algorithm finds a path with the lowest total error through the matrix starting at (1,1) and ending at (I, J). The algorithm accumulates the error across all the $max(i, j)$ pairings. The implied hypothesis about temporal regularity of speech is that speech production can change tempo so rapidly that it can be changed by an arbitrary 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 that, just as the linguists insist, linguistic information, such as word identity, lies in *the sequential order of the states.* 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 timing as a kind of noise* to be ignored in this clever way.

Phoneticians have never been happy about this way of treating time. There is too much phonetic evidence that subtle details of speech timing carries usable information about word identities [Klatt, 1976,Port and Anderson, 1989]. Instead, they have preferred to assume that listeners (like

[4]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 [Port et al., 1988] even though these authors did not claim that their technique would be generally practical.

phoneticians themselves) have available to them a time window on the signal, a stretch of the signal that is available at one time. This class of models will be discussed next.

3.2 Static Full Bandwidth Windows

The third model assumes that more powerful information is available in the signal, but it is conceptually similar to the second since it simply uses the time buffer (whether the buffering is done mechanically on audio tape or digitally in a computer file) as a 'full-powered' time window. That is, the system simply preserves the entire bandwidth of the input signal and stores it. The system (or researcher) then extracts whatever temporal measures look like they might be useful. Neither phoneticians, speech recognition engineers nor psychologists modelling human perception hesitate to measure units such as 'vowel duration' or 'voice-onset time' for the purposes of their models of perception. But these measurements require that there be a display in which time has been converted to physical distance. The undeniable convenience of window displays for signal description and for mathematical models of perception, however, contributes nothing to their biologically plausibility. Since there are so many variants of the time window, I will mention several examples. The first is simply the sound spectrogram.

3.2.1 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 by intensity. We can collect evidence this way about the kind of information that the perceptual system must be using. If perception experiments are also done, it can be shown that many of the temporal properties of natural speech are detected and used by listeners. By and large, most prominent temporal regularities can be shown to play a role in speech perception if a sensitive experimental task is designed (see [Klatt, 1976]). Additionally, we need a theory of how the temporal patterns we observe could be extracted from continuous signals. In universal practice for analysis of speech, information is obtained by turning time into physical space, either in a data file on a computer disk or on a piece of spectrograph paper. Since a timer that measures particular physical events cannot be built until one knows exactly what one wants to measure, the most natural way to implement time measurement is as length along a window containing the sampled waveform. The sound spectrogram in Figure 2 below has a distance measure for time. Most speech recognition systems have a window in which time intervals can simply be measured as distances along the sampled waveform (or along one of the integrated signals derived from it). Speech scientists take temporal measurements from such displays (see, for example, [Lisker and Abramson, 1964,Klatt, 1976,Port, 1981]) without being concerned with how a brain might achieve this. Although the notion of 'echoic memory' has been proposed [Neisser, 1967], there is no strong evidence that this representation is actually the raw acoustic signal. After all, the real echoes we hear in canyons blurr into white noise unless a brief sound is preceded and followed by relative silence.

Use of a window on the input signal has some consequences that make it implausible for a biological system. How long should the buffer be? Optimum buffer length follows from the kind of information that needs to be extracted from the signal. The window size must be large enough in time and in number of frames to capture useful features of the signal. There is evidence of a rich auditory code (see [Hawkins and Presson, 1986] for a review), for a duration between 1/2 sec. and 2 sec. This representation captures very detailed information about the signal. But does it, in fact,

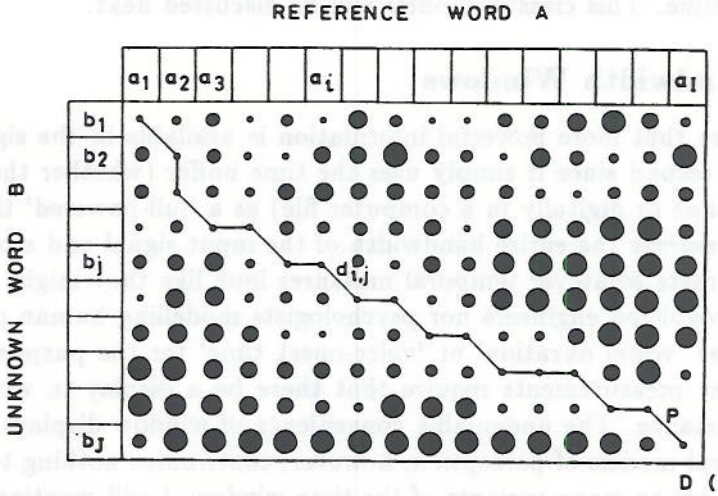

Figure 1: Dynamic time warping is done by finding a minimum error mapping from the reference (or template) vector to the test item vector. The vector for each time slice is subtracted from all other vectors. The magnitude of those differences is displayed as circles. The algorithm finds a path through the matrix starting at (1,1) and ending at (I, J) and accumulates the error across all the $max(i, j)$ pairings. This error is used for comparing matches from different words.

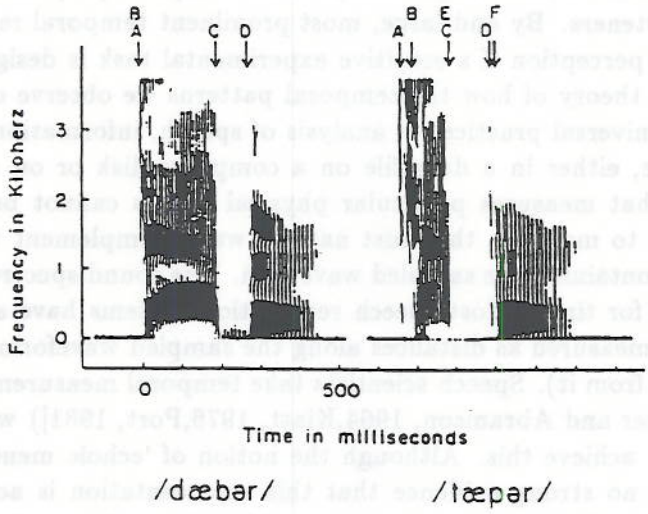

Figure 2: Spectrograms of the words *dabber* and *tapper* showing A, the point of the stop burst, B, the point of the voicing onset, C, the onset of closure for the medial stop, D, release of the medial stop, E, voicing offset for the medial stop, and, F, voicing onset for the medial stop. It can be seen that the two segmental differences between the words are manifested as a rich set of durational differences. Measurement of these points is very difficult to do automatically. Manipulation of parameters like these in synthetic stimuli in perception experiments shows that they are used by listeners to natural languages. But does a listener have a spectrogram to take measurements on?

save the entire bandwidth of the signal? Probably not, as I will argue in a later section. Meanwhile, there is another full bandwidth static window that is very familiar to workers in cognitive science.

3.2.2 Hearsay2

For speech recognition systems, windows are the easiest way to approach the problem of temporally distributed information. Thus, the Hearsay2 model [Victor R. Lesser and Reddy, 1975] processed a whole sentence at a time by placing it in a buffer. A huge data structure called the 'blackboard' representing the whole utterance was defined. It records the various kinds of information, including the raw waveform as only one, that are extracted from the signal. Each kind of description is stored at the appropriate point in time. This structure permits processing of information that climb up from more local to more global generalizations and also to work downward, from high-level descriptive information downward to reinterpret lower levels. Despite the intuitive appeal of the bidirectional reinforcement, it must observed that a static blackboard that allows access to all levels of an entire sentence at once is extremely implausible as a model of auditory perception. How could sentence beginnings and endings be found in advance?

3.2.3 Elman-Zipser Nonoverlapping Windows

Spectrograms are just displays while Hearsay2 analyzed full sentences from a restricted grammar. Time warping also uses a buffer containing a fixed-length stretch of speech. Within the connectionist framework, many systems for speech recognition have also employed windows that are long and more or less nonoverlapping. These present the network with the entire bandwidth of enough signal to allow the pattern to be displayed. The basic idea is to have a separate set of nodes to represent inputs for each time slice. One recent example is the feedforward network trained with back-propagation in Section 3 of Elman & Zipser (1988) . This system 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, 3 hidden nodes and 9 output nodes. The system was trained to label what is found in the window. Elman and Zipser showed that this sort of network successfully categorizes stop-vowel syllables despite variation across hundreds of tokens. The network appears to capture phonetically useful information about temporal location by converting time into a spatial representation within the window. But is this really spatial? On closer inspection, from the standpoint of the network, the nodes representing each time slice are just input channels. They do not have 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' does not arise). The window duration must be fixed since window size determines network size. Obviously, animal nervous systems do not have static buffers resembling this.

As a theory of cognition the static window solution, with or without labels indicating serial position within the window, has little to support it. There is no evidence of strobe-like effects due to interference with the periodic 'blanking interval' the time point where window contents are replaced or where attention switches from one buffer to another. 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 has been obtained, it must be assumed that this window has either

a fixed duration or a stochastically varied window duration (that would help avoid stroboscopic interference effects). Nothing is known about the signal until after the window is full. Only then can some pattern be used to control processing.

Of course, phoneticians have certainly never claimed that the $f \times I \times t$ spectrogram of 1.2 sec duration is a part of "the theory of speech perception" or "the theory of audition," but reflection shows that something like it must implicitly be postulated if parametric measurement of time intervals is to be performed by 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 they convert time into place, such models explicitly separate the representation of time from perceptual analysis of time. But parsing of the input in this way, requires breaking experience of the world up into temporal windows, that is, into periodic alternations of information collection with information processing. An assumption of this kind of temporal multiplexing is that the window must be long to recognize long patterns. But in order to be long enough for the longer patterns, the ability to respond to short patterns is artificially delayed – even for patterns that are important and require reaction time that is as fast as possible. The system in this model must wait for a time that could be as long as the window duration plus the time needed to switch attention from the old buffer to the new before the analysis can even begin. This technique makes the fastest possible reaction time dependent on the longest duration patterns the system needs to recognize. Clearly, the criterion of a quick response is too important to animals. I conclude that long patterns must to be extracted by some technique that permits streaming.

Whereas 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 possible to have varying amounts of overlap up to the maximum at which the window simply slides over a single spectral frame with each time step.

3.2.4 Overlapping Windows

If windows overlap maximally in time, that is, if each input frame moves successively through each position in the window, from first position (most recent) to the last, then the system will 'stream', that is, you could run it continuously. The familiar NET-talk system [Sejnowski and Rosenberg, 1986] for translating orthographically spelled text into phonemic transcriptions is an example. 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 and 3 backward across the input. These connections across time are also sometimes called delay lines.

Delay Lines This variant of the window technique is currently under intensive development within the connectionist speech recognition community. These models use delay lines (as in the simulation in Section 6 of [Elman and Zipser, 1988] and in [Waibel et al., 1988,Watrous, 1990]). Here events at several adjacent points in time are presented simultaneously to a single node in the network. This is like a window since physical places (that is, input lines) represent distinct points in time. Nodes with delay lines simply sum the inputs from different points in time. If only a small number of delays are sufficient, they can supply useful information about the history of the signal

without raising too much of a problem of shift invariance. Sensitivity to temporally 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, this system can run continuously, with each input marching down the row of delays and 'falling off the end'. Such a design implies that learning a temporal template should come most naturally to it. For example, if VOT tends to be 50 ms, then the system should learn to look for a configuration of particular events at t_0 and at t_{-50ms}. The learning should be fairly sensitive to changes in the time scale. That is, if VOT should appear infrequently at 25 ms or 100 ms, then the system should have difficulty being responsive. Still, delay lines are a plausible wiring technique for a nervous system that needs to analyze temporally distributed patterns.

Thus far, the window systems considered save the input signal itself. That is, one way or another, they save the full bandwidth of the input. No selection process is applied to the signal in the spectral domain. Analysis of spectral contents does not occur until after the window is filled. Another large class of models include subnits that select certain kinds of spectral information.

3.3 Selected Feature Windows

Some models for temporal pattern recognition select portions of the frequency space and track particular kinds of events. They do so 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. For example, Kewley-Port [Kewley-Port, 1983] has suggested that to identify the 6 syllable-initial stops, information events distributed over time must be tracked (in contrast to [Jakobson et al., 1952] and [Stevens and Blumstein, 1978]). Another example, one that implies a more general approach to temporal pattern perception, is the slope-detector model of Smythe.

3.3.1 Slope Detectors

One additional method for extracting information about temporal events from a time window is to have hard-wired detectors of change in the signal over time.. Activation of certain subnetworks indicates that specific event sequences took place. This approach assumes a bank of detectors in order to span the acoustic space. These are essentially dynamic feature detectors. The idea of neural modules that are specialized for particular functions is surely a likely possibility for human perception. The difficulty is to determine what particular functions are actually modularized. The representation of an event in this model, then, is increased activity in a particular node or set of nodes. 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. Thus the instantaneous state of the network can carry information both about what happened in the past as well as when it occurred.

Obviously, this is a very large class of information gathering techniques, so a single concrete example will be given. Recognition, prediction, decision-making and memory tend to become enmeshed and inseparable in these systems. One cannot easily tell just where representation leaves off and decision-making begins. The input arrives and, if it has the right properties, that is, if a peak occurs in a particular frequency region, then activation is accumulated over time in a summing device. Otherwise, no activations are accumulated.

Study of neural structures in the rabbits visual system ([H. B. Barlow, 1965]) reveals banks of frequency×time detectors to measure the rate of change in some part of the signal. This idea has been explored by Smythe[Smythe, 1987,Smythe, 1988] and by Watrous [Watrous, 1990] for recognizing schematized formant transitions. Smythe's system used a handwired set of mini-networks for identifying the slopes of linear formant tracks. As shown in Figure 3A, the model has a battery of detectors for various positive and negative formant slopes across the frequency range. It is designed so that competition will result in only one of these states being an attractor for a given input.

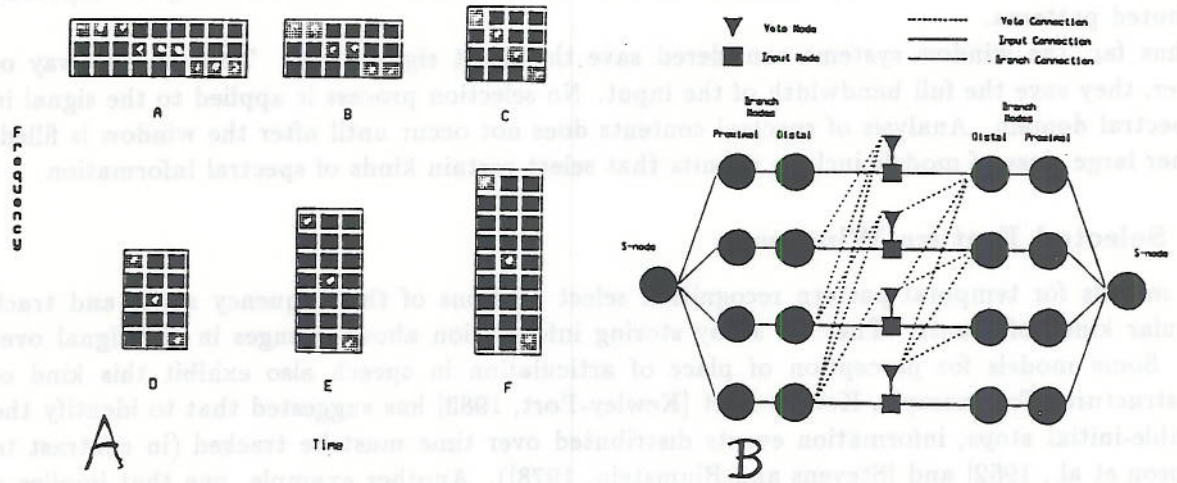

Figure 3: **A.** The input fields for several motion detectors with different slopes. Each column represents a 5 ms time slice. A lightly shaded square indicates the presence of a formant center at the frequency corresponding to that row and time slice. A motion detector is tuned to identify one of these transitions. **B.** Each slope detector is a 'veto inhibition network'. It permits strong activation if its inputs are excited in one direction but is suppressed, or vetoed, if the excitation occurs in the wrong direction. In B is shown a pair of detectors sensitive to slopes with opposite sign. Two motion detectors for a rising or falling transition of the same rate. The inputs arrive at the square nodes in the middle of the figure. The detector on the right is tuned to the falling transition and the detector on the left is tuned to the rising transition. Dotted lines are connections that prevent the next node (as a circle) from firing for a short period of time.

The inputs are shown in the middle of the display and detector output nodes (or S-nodes) for each positive and negative slope is shown at each side. The dotted lines represent inhibitory 'veto edges'. It can be seen than any input, passing across the series, from either the top or from the bottom will allow one S-node to collect activations from all the inputs, while the other detector node will get activation from only one input node. The slope detectors inhibit their neighbors so that the most active slope detector nodes in each frequency range are stored in a line of delay nodes.

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. These particular slope detectors are rather inflexible.

3.3.2 Iterative Nets

Another way to move toward a dynamically functioning net without having recurrent edges (since backpropagation works only for feedforward architectures)unwrapped network is roughly equivalent to a much smaller recurrent network. This variant of the time window has inputs that arrive at different places in the net [Hinton, 1988,Elman and McClelland, 1986b]. The reason it is grouped with selected feature models is that at each time slice after the first, what is retained about earlier events is an analysis of those events, not the original signal itself. Thus, these models are more sophisticated and plausible as psychological models since they exhibit dynamic, instantaneous response to acoustic inputs. They still have the problem that at the close of stimulus presentation, they must be reset to an initial activation state for presentation of the next stimulus. There is no way to make TRACE run continuously.

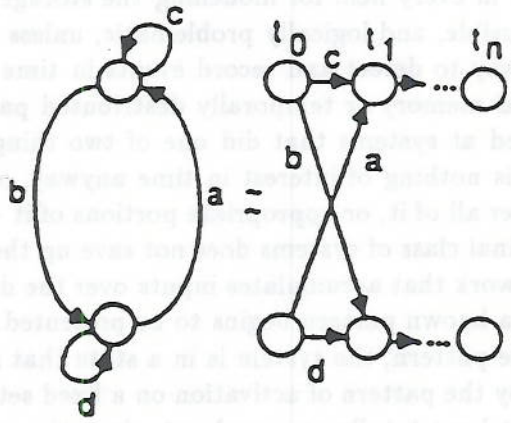

Figure 4: A simple iterative network in which stimulation for different time frames are supplied to different input nodes. A single kernel network at the left is basically repeated for each time slice. The flow of activations through the network simulates recurrent activiyy through time. Afer each trial, the system resets all activations and the next window begins again at t_0. One advantage is the the architecture remains feedforward so that backpropagation can be used.

In iterative systems, sequential points in time during presentation of a trial have distinct input nodes. The architecture for the purposes of learning is feedforward, but at least from the onset of stimulus presentation until the end of the stimulus, it behaves like a kind of dynamic network. At the end of each trial, the system is externally reset to t_0. So, the maximum duration pattern that can be detected is constrained by the hardware of the network.

Are Windows Plausible? In general, static windows have the advantage that lengthy patterns in time can be represented all at once so that collecting the information is all done for you by the window. Different points in time are located at physically different points in space. The pattern recognition system can then look for patterns within the window. Having a single long window has the disadvantage, however, of the problem of *shift invariance* [Tank and Hopfield, 1987,Hinton, 1987,Waibel et al. How does the system know that something in one position in the window is the same as something somewhere else in the window? The longer the window the more serious is this problem. This problem exists because it is not the pattern that determines the relative size of the window, but the hardwired network structure. Short windows for special purpose features, however, are restricted to a fixed set of 'hardwired' pattern features.

Windows as input buffers are useful for converting time into static location and for processing input from various points in time. All the various techniques discussed in this section are fundamentally very similar. Each receives input either simultaneously or sequentially within a trial and must then be reset to an initial state before another input can be processed. But the shift invariance problem – that a pattern located in one place in time is distinct from a pattern located in a different place – is created by the fact that patterns may come in different lengths: some fill the entire window, others do not. Because different time locations are handled by different weights in the network, some temporal objects must recognized in multiple locations [Hinton, 1987]. Although windows have been popular in every field for modelling the storage of information over time, they are biologically quite implausible, and logically problematic, unless they are very short and stream their inputs. Some better way to detect and record events in time must be found for biologically plausible models of temporal memory or temporally destributed pattern recognition.

Thus far, we have looked at systems that did one of two things. They either insist, like the ordinal models, that there is nothing of interest in time anyway, or they create a window on the signal by saving it up – either all of it, or appropriate portions of it – so that useful patterns can be extracted all at once. Our final class of systems does not save up the signal itself nor does it have a hardwired circuit or subnetwork that accumulates inputs over the duration of the pattern. Rather, it enters a trajectory when a known pattern begins to be presented. Events in the past change the state now. At the end of the pattern, the system is in a state that represents pattern identity, not by where it is (that is, not by the pattern of activation on a fixed set of nodes), but by the sequence of activations. Time has not been totally squeezed out, since the representation still requires time to manifest itself. And the duration of activity by the pattern analysis unit is not fixed apriori. Instead, when trained to know something, the system learns the appropriate size of the descriptive window for pattern units. The 'identification state' can only be reached by following a particular trajectory that is learned.

3.4 Dynamic Attractors

The fourth and last type of model considered in this essay 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, but there is some behavioral evidence 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.

3.4.1 Elman's BI-DAA-GUUU Network.

One simple system that runs continuously and stores stimulus history internally was explored by [Elman, 1988]. He proposed a simple recurrent connectionist network for tackling an interesting problem in sequence, as shown in Figure 5. The values of hidden nodes are fed as input to a set of context nodes. Each node in the hidden layer sums inputs from outside and from the context nodes. He constructed many random strings of BI, DAA and GUUU coded into 6-bit vectors (or 'distinctive features') presented one character at a time. After optimization of weights on the feedforward edges, the system learns to predict most of the predictable properties of those strings. Its average error for the dimension that distinguishes the consonants from the vowels is high in just those positions and very low elsewhere. 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 observed.

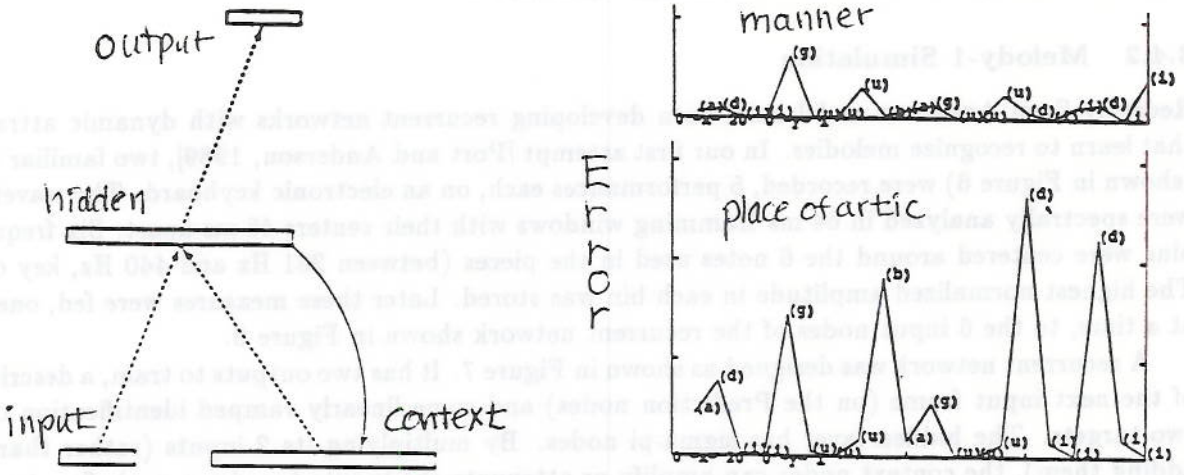

Figure 5: On the left, is Elman's recurrent network for the BI-DAA-GUUU problem. The input is fed one slice at a time. The hidden context node stores information about recent inputs. As shown on the right, the system can predict the bit corresponding to the place of articulation feature of the consonants very well – with a periodicity corresponding to the pattern of inputs. The feature distinguishing consonants from vowels is always predictable.

Prediction is an essential task for networks that process variable input. Success at this task allows us to employ a target-driven learning algorithm like back propagation. This 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. Real nervous systems cannot exploit such a simple clock.

Recently, there have been numerous attempts to construct systems that can learn to follow dynamic trajectories that differ depending on what input patterns arrive. These dynamic systems (a few examples are [Sompolinsky and Kanter, 1986,Keeler, 1988,Bell, 1988,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 two dimensions of the model are plotted against energy [Smolensky, 1986], then we can conceptualize the instantaneous state of the system during a recognition problem as a ball rolling downhill in this

3-D 'landscape'. The system is moving toward a 'low energy state' (or, in information terms, a state of greater harmony). In the systems described above, the goal of design and training was to assure that the system would find the most harmonious state (analogous to a least-energy state for a physical system). Hopefully, that state would 'represent' the correct answer as an attractive fixed point of the system. On each trial (after training), the system falls to one of the target states. It is reset before the next trial. In fully dynamical systems, however, we want to assure that fixed points are never reached. Once at a fixed point, the system is completely still. We want attractors, but the system should only approach them. That way, depending on stimulus input, it can move off elsewhere if necessary. It can be guided in time by the input itself. This implies networks with heavy reliance on recurrent edges. It also seems clear that higher order relations between nodes is required, that is, activation products in order to assure rapid response. In our nets the hidden nodes are sigma-pi nodes that allow the input and memory information to gate each other. In this way, successful prediction can be noted and made use of.

3.4.2 Melody-1 Simulation

Recently Sven Anderson and I have been developing recurrent networks with dynamic attractors that learn to recognize melodies. In our first attempt [Port and Anderson, 1989], two familiar tunes (shown in Figure 6) were recorded, 5 performances each, on an electronic keyboard. 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 highest normalized amplitude in each bin was stored. Later these measures were fed, one slice at a time, to the 6 input nodes of the recurrent network shown in Figure 6.

A recurrent network was designed as shown in Figure 7. It has two outputs to train, a description of the next input frame (on the Prediction nodes) and some 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 amplify or attenuate the inputs (or vice versa). So, since our measures were always the same length, the system automatically resets itself at the end of a trial.

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 (as can be seen in Figure 6). Of the 5 recorded versions of each measure, 2 were used as training tokens and 3 performances were reserved for testing. The competing 'noise' measures were the 14 other musically distinct measures. The measures were presented continuously.

Networks of this architecture cannot be trained with backpropagation (since recurrent edges also were trained), so the Williams-Zipser algorithm ([Williams and Zipser, 1988] 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 3.0 [Swets, 1961].

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. It seemed that this might be due to a weak representation of durational information. In order to allow a better correlate of duration measurement, the next network was designed. The idea was to use

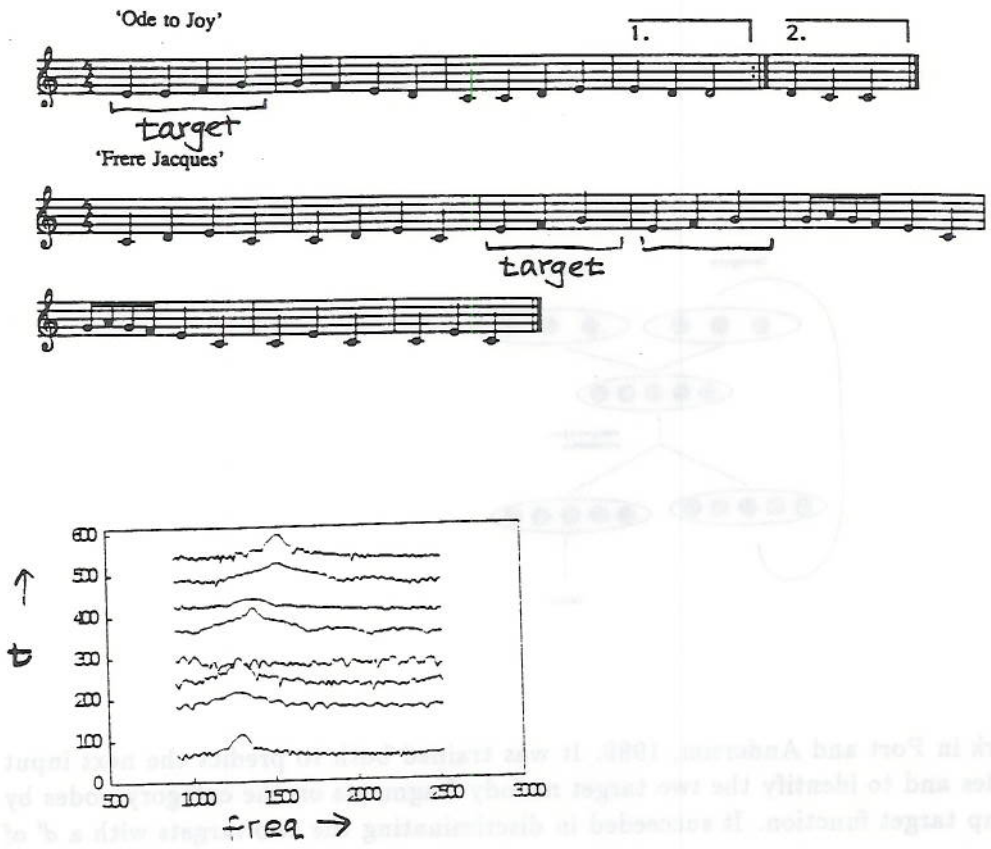

Figure 6: Stimulus construction for the melody recognition task. Panel A shows the original music with the two target measures marked. Panel B shows examples of the FFTs (fast Fourier spectra) for one measure.

sigma-pi nodes to control the activation level of some nodes so that they would measure durations.

3.4.3 Melody-2 Simulation

In our second simulation of melody recognition, task, we have networks with a fully recurrent Dynamic Memory, as shown in Figure 8 (see [Anderson and Port, 1990]).

All edges in the figure had learnable weights. The Acoustic Feature Units represent the notes of the scale and their recurrent edges provide a bit of reverberation to add acoustic 'fill' to the original staccato performance. The nodes of the Duration Sensitive Module were designed to produce a peak output activations that are proportional to the duration of their input. In order to do this, they must shut off output when the input feature ceases. This is achieved by having the activation of each Acoustic Feature node serve as a gate for its Duration Node in the form of a sigma-pi connection. The Duration Nodes are linear, so that as long as they continue to receive input, their recursive gain pushes output asymptotically toward a fixed point. This can be seen in Figure 9, which displays the outputs of the Feature Nodes, the Duration Nodes and the 2 Category Nodes. Longer duration

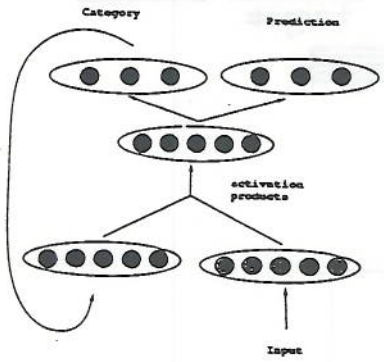

Figure 7: The network in Port and Anderson, 1989. It was trained both to predict the next input on the prediction nodes and to identify the two target melody fragments on the category nodes by following a linear ramp target function. It succeeded in discriminating the two targets with a d' of better than 3.5.

inputs in the Feature nodes produce higher peak activations for the Duration Nodes. When the Feature node outputs a 0, it zeros out the recurrent link on its Duration node, returning it to 0 on the next cycle. The two Category Nodes, one for each target measure, were trained to produce a linear ramp whenever their measure was presented. Again, the Williams-Zipser learning algorithm was employed to optimize weights even on recurrent edges [Williams and Zipser, 1988].

This system learned to recognize the melodies better than Melody-1, getting 87% and 89% correct for the two target measures. The most interesting part is the dynamic behavior of the short-term memory. To see better how these 7 nodes acted, we performed *principle components analysis*, which rotates the 7 dimensions to find a small number of dimensions that exhibit the most variance. Thus Figure 10, shows the first two principle components for continuous presentation of several target measures and for several nontarget measures. It can be seen that for all the nontarget measures these nodes cycled around one region of its statespace. Since we see tracks here, we may imagine an energy landscape lying across the two principle component axes. When a target measure begins to be presented, the system moves through a very characteristic sequence of states as it falls through an energy landscape that is partly created by the stimulus input sequence.

So 'recognition' of a trained pattern is exhibited in this short-term memory not by a single node 'lighting up' (which is what the Category Nodes were trained to do). Instead, the memory tracks the pattern by moving through its state space. It seems that one effect of the training procedure

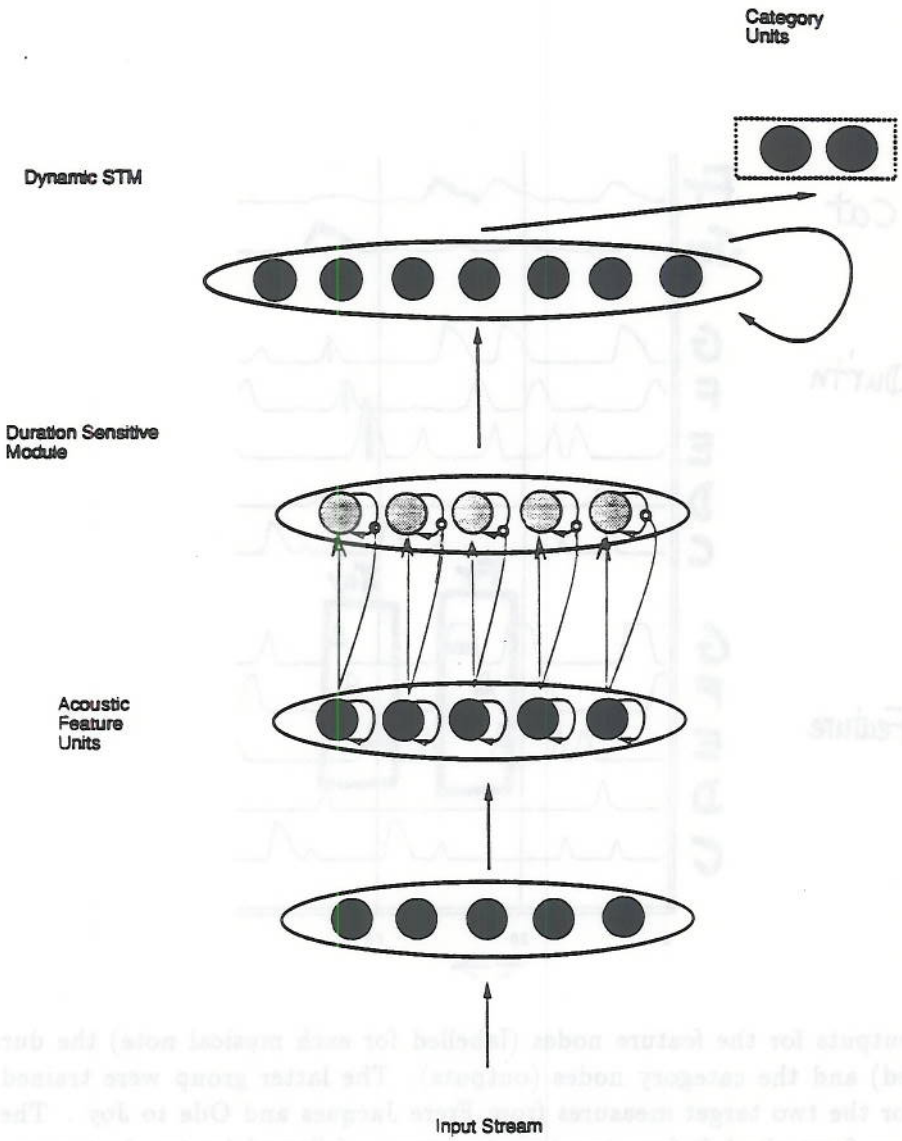

Figure 8: Dynamic network used for the melody recognition task in our second model in Anderson and Port, 1990. There are 6 Input nodes fully connected to 6 self-recurrent Acoustic Feature nodes which are fully connected to 6 Dynamic nodes. These produce linear output. The dynamic memory nodes are fully connected to each other.

on this network architecture was to construct an energy landscape that can be restructured by the incoming stimuli themselves.

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 11, 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 statespace – only 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.

Thus, the system is able to recognize patterns distributed in time even under certain distortions of time. It was still able to differentiate patterns that differ primarily in durational structure. It does so by recursively exciting itself such that, as long as inputs continue to support the pattern, then it follows a trajectory through its statespace. This trajectory was followed despite major distortions of duration. Presumably, it will also exhibit resistance to other kinds of 'noise'. We

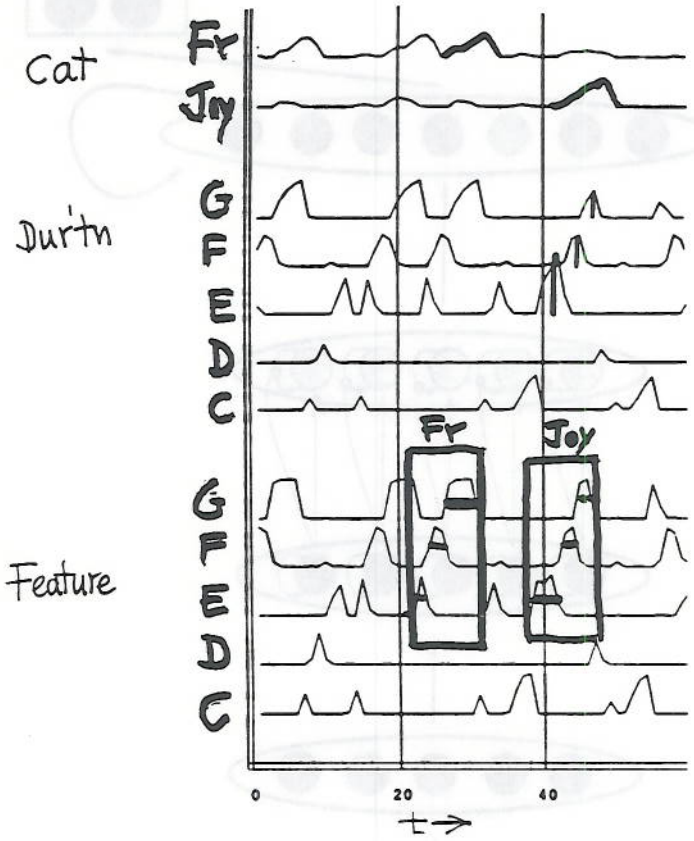

Figure 9: The outputs for the feature nodes (labelled for each musical note) the duration nodes (similarly marked) and the category nodes (outputs). The latter group were trained to a linear ramp function for the two target measures from Frere Jacques and Ode to Joy . The continuous outputs are shown from the left for a nontarget measure, followed by two instances of the *Frere Jacques* target, one nontarget measure and then the *Ode to Joy* target measure. Finally another nontarget. One token of each is boxed at the feature nodes.

plan further exploration of the resilience of this kind of system.

This kind of representation appears to have certain clear advantages for a nervous system. First, it has a built-in mechanism for reset when a pattern (of whatever size) is completed. That is, it should be able to learn patterns of any duration. The outputs shown above plus results of other unpublished simulations show that sigma-pi nodes are very effective at allowing inputs to control the processing itself. Second, response can be initiated as soon as possible, that is, as soon as the information in the stimulus permits a decision to be made. Third, 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. Finally, the system reaches recognition without having to label or identify any components. It only gets a 'label' for the entire sequence. That is, in a sense it has employed a 'subsymbolic' representation of temporal pattern to enable symbol-like recognition of each pattern as a whole.

The model has other implications, however. One property is that scaleup of this idea seems to imply learning a large number of state trajectories for complex auditory patterns of various sizes and complexity. One key implication of this model for a psychological theory of time is that there is nothing that represents time. There is no 'chronotopic field' such that distances along it represent

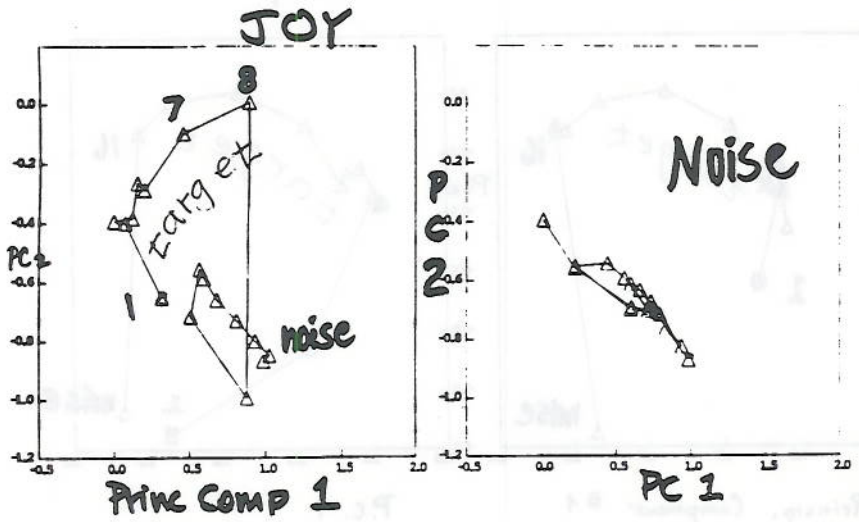

Figure 10: The first two principle components for the dynamic short-term memory node activations. Each triangle represents the state after one input frame. There were about 8 frames for each target measure and they are connected in the order in which they were presented. The loops near the center of each frame correspond to nontarget measures, while the arch toward the top in the upper left frame is the dynamic trajectory of eight frames of the *Ode to Joy* target measure.

intervals of real time. The way in which the states should be made to change to represent useful properties of the signal may be difficult to work out, and threatens to be adhoc for each perceptual task (that is, for each pattern at this level of representation). But this multiplicity may be the right kind of multiplicity. A learning process is clearly required to guide development of a model for the states themselves. On the other hand, as will be argued in the next section, there is evidence that human perception may depend on a similar kinds of learning.

4 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 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 (see [Hawkins and Presson, 1986] for a review). In these experiments subjects typically listen to stimuli that are either very simple (eg, pure tones) or else very familiar (spoken words). The important question, however, is whether it is an *acoustic* store. Is there a true *chronotopic field* containing spectra displayed through time which subjects 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 (see [Watson and Foyle, 1985,Watson and Kelly, 1981]). 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.

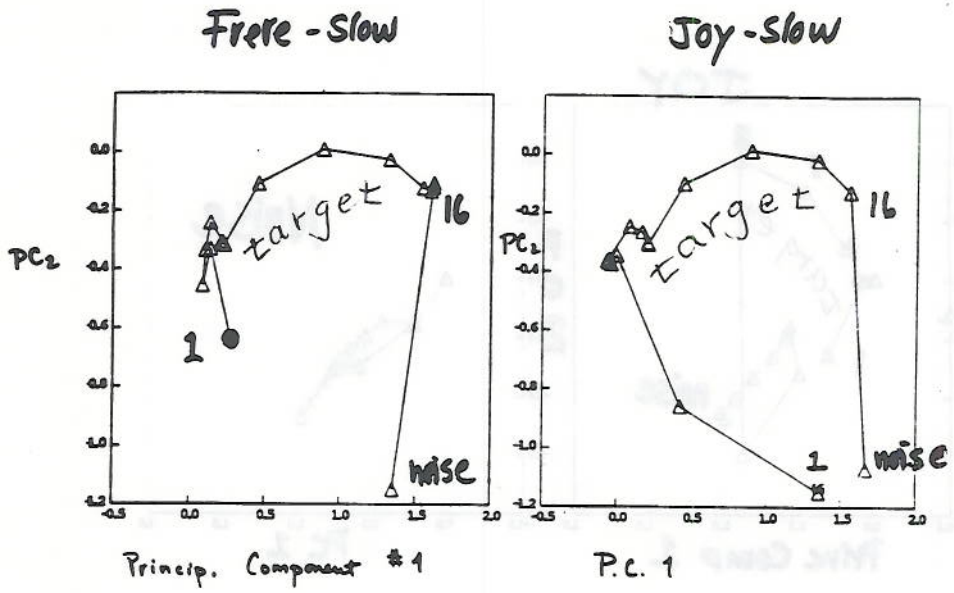

Figure 11: Same display as the previous figure except that the stimuli were presented to the network at half-tempo (by repeating each spectral frame twice). The same trajectory was followed for the targets as before, but it now takes 16 frames to complete the trajectory instead of 8. Each window ends at the beginning of a noise measure. No retraining was required for this performance.

The basic idea of this research program was to make very complex acoustic patterns and to evaluate subjects ability to make detailed judgments about the patterns. That is, in terms of cognitive science, they ask about the quality of auditory representation. Typically subjects were presented two very similar sequences and asked if they are the same or different. 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 (around 70% correct). 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 complex burst of sound that perhaps resembles a turkey gobble and lasts less than a half a second. The subject is played one of these sequences twice and asked if the two are identical. The second pattern may have one of the components 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 all these parameters are, of course, well know 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 do listeners have access to? Can this representaion be improved with practice?

The general finding is that if the random patterns presented to a subject in a single sitting are drawn from a large set (eg, more than 20 patterns), then subjects are unable to do this task very well. Even with thousands of trials drawn from a set of 20, they require enormous changes in stimulus parameters for the difference to be detectible [Watson and Foyle, 1985]. For example, in many cases frequency must be changed by several octaves before subjects can detect the change reliably. The 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 a single pattern at a time (so that he

has little uncertainty about the pattern that will be presented), then his performance at detection of a frequency change gradually approaches his performance on the tones if they were presented under ideal experimental circumstances (which presumably represents the psychophysical limits). However, to approach asymptotic performance still takes several thousand trials on each individual pattern [Leek and Watson, 1984]. It is possible that the training involved here produces a similar representation to that achieved by the [Anderson and Port, 1990] network described above. That is, training may produce a consistent dynamic trajectory through state space for a learned pattern.

We speculate that initially subjects had no way to represent the complex pattern so they were forced to integrate the entire sequence across time. This may be what they are doing in the case where there is high-uncertainty about what pattern will occur next. Successful discrimination in this situation may depend primarily on an integrated superposition of the sequentially presented tonal components. As a result of extensive training, however, in circumstances where they can focus on learning a very small number of patterns (especially if it is only a single pattern), the subjects can learn to follow a distinct dynamic trajectory for each component that enables them to 'hear out' small changes in individual components.

There are several kinds of evidence that such an account may be on the right track. The first is that it appears that when subjects listen to patterns, correction for tempo variation is obtained quite naturally, almost 'for free'. 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]. When the task is thought of as an information processing task, then the tempo invariance makes sense since the information is the same although at different rates (as pointed out by Watson et al). 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 (very slowly) dynamic trajectories through state space. As the pattern or pattern component presents itself, the system is driven through this trajectory. Some deviations in the pattern, such as changes in the overall rate of presentation (or masking by patterns in other frequency ranges), should not disturb it very much.

Thus, these behavioral results resemble our dynamic memory system. Both require extensive training and are resistant to backward masking (since patterns are recognized *as they come in*, not after a period of subsequent processing). Because of their dynamic behavior, the temporally distributed patterns are resistant to either spectral or durational noise. Changes in tempo that are constant across presentation of the pattern cause little or no decrement in performance.

To pursue this parallel further and to directly compare a model using time delays with a dynamic model will require careful experimental comparisons. For example, if delay lines play an important role in perception of auditory patterns, then recognition of patterns that depend on time differences in absolute value (eg, a lag of x ms) should be easily learned (since the weights at a particular lag could learn to be strong and all others weak). Similarly, a time-delay system should be able to construct a 'temporal mask' and pay attention only to certain events at fixed distance, totally ignoring events at other temporal lags. Watson's results suggest that a temporal mask, especially one specified in absolute values, is virtually impossible for humans to learn. On the other hand, patterns that are defined in tempo-invariant terms (like a melody or word produced at a range of tempos) should be difficult for a delay-line system. A dynamic model should have a difficult time processing events at a fixed time lag so that intermediate events are ignored. The time normalization and trajectory tracking of this type of system should make performance on such

tasks very poor. By means of experiments directly comparing the limitations of various network models and human behavioral data, the nature of short-term processing of auditory patterns should be systematically explored.

5 Comparison of Representations

Natural languages use patterns in speech signals that extend over lengthy intervals. These patterns are employed to help distinguish words. Human listeners can apparently use subtle distortions of speech timing to extract useful information (see [Port and Crawford, 1989]). Human perceptual systems can evaluate many kinds of rhythm over a wide range of periods. This implies memory that is rich in temporal information. Perceptual evidence from human listeners shows their ability to make use of such patterns.

This temporal richness might conceivably be achieved by a tape-loop kind of storage in which the earliest information is dropped off the end. Delay lines are an implementation of this effect. Because they store the acoustic signal at a bandwidth equal to that of the input, they are 'expensive' – both for computers and for nervous systems. Although useful for some purposes and probably used in nervous systems for some purposes, they probably do not offer an adequate model for the recognition of temporal patterns.

Although I have not claimed that this survey provided a typology of models, there are some typological properties that can be raised in comparison of these models. I find at least 5 issues by which they may be distinguished. Many of these have been alluded to throughout the text.

What is the spectral representation like? Some models save the entire input spectrum, that is the entire input frame (no matter now many bits are involved). For example, Hearsay2, Elman-Zipser and Sejnowski-Rosenbaum save the entire input signal itself. Other models, like the Smythe system have a large number of parallel windows that each select for particular frequency ranges. Other systems need not have their spectral description selected in advance, but can learn what to pay attention to in the spectrum.

What is the duration of the window? One possibility is to have the duration of the window as short as possible. It can approach the continuous case for delay-line models or our own Melody-2. At the other extreme, sometimes windows are proposed that are as long as the longest relevant patterns – as in Hearsay2. If something in between is proposed, then both constructive processes (to build longer patterns) and analytical processes are required.

How are the detector systems reset? If pattern recognition takes time, then the system is 'tied up' tracking the pattern. Eventually, the system must be reset to some rest or background state as various possibilities eventually fail. The issue here is control over processing. How is this achieved? This is a fundamental issue. For Hearsay2, reset comes at the close of analysis of an entire sentence, something which is determined external to the entire model. Iterative networks like TRACE are also automatically reset by the end of the trial (as well as the end of the network). The Elman-Zipser model had carefully edited speech which led to reset at the end. The Elman BI-DAA-GUU network, however, is more sophisticated. It has no external reset but learned from the sequential distribution of the input patterns when a new component pattern would begin. These

points, where it 'expected a consonant' are effectively reset points. Similarly, our melody networks were automatically reset by the input patterns themselves. After reset, they returned to a waiting state. In this way, control of processing is guided by the input pattern itself. The disadvantage is that learning is required, and learning is expensive.

Do the minimal units have 'meaning'? At one extreme, Jakobsonian distinctive features are linguistically meaningful and were supposed to be defined directly in the acoustic signal. The minimal units of a slope detector system also have a clear meaning, though further processing is required to determine the identification of stops. At the other extreme, the internal states for a dynamic model, like Melody-2, have very little meaning in themselves. They are representations that are fully distributed, not only in space, but also in time. They acquire meaning for a pattern only in a specific context.

Is the time scale of detector output ordinal or rationale or something in between? In Jakobsonian distinctive feature theory as well as in dynamic time-warping models, the output is only ordinal. It seems that words, as abstract objects, really do have an ordinal structure that is well-captured by ordered phoneme-like or syllable-like units. But as acoustic and articulatory objects, they clearly have a far more complex temporal structure. That is, words as pronounced are *not* invariant under all the transformations that are required by ordinal time. Nervous systems need ways to get much more information out of temporal pattern. But the only other alternative seems to be a raw time window (as found on sound spectrograms and on the blackboard of Hearsay2). This commits one to using milliseconds measured on a spectrogram as a model for time. How can a model that lies between these extremes be defined? Dynamic models just may be flexible enough to learn a variety of invariance transformations appropriate for human speech.

These five questions help provide some perspective on the ways in which models of temporal representation differ from each other.

6 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 and experience has a representation. A clear distinction between operations and representations may be impossible to make in general.

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 seems almost inescapable that scalar properties of time can either be ignored and processed as serial order, or else, if that can be shown to fail, that time can be treated simply as another parameter. But if it is a parameter, then it must be measured. And this task places a huge burden on some preprocessing system that must both display time as space and then recognize certain landmarks and measure the required values. If our concern is simply with the description of sound, then scientists may just go right ahead measuring. But if such measurements are supposed to play a role in a model of cognition, then we are responsible to account for the extraction of this information. No viable models appear to exist currently. What I have tried to show in this paper is the very narrow range of underlying models that have actually been exploited in the literature of the many disciplines with a concern for speech. In addition, I have tried to show that another model is possible, one that is intrinsically dynamic.

As for the basic issue in cognitive science of how the central nervous system functions, these results on timing have broad significance. It is quite clear that the time-integrating feature detectors so favored by linguists cannot capture the information listeners use. And time-warping models discard critical information. These results imply that the mind cannot be ruled solely by principles of sequence and order (as logic and mathematics are). Instead, the brain is a dynamic system, something constantly in a flux of activity. There cannot be discrete symbols that could be put into 'buffers' and expected to stay there until you do something to erase them. Events in time, at least, probably cannot be put into such buffers. At the lowest level, the firing patterns of the nervous system have to interact dynamically with incoming signals. Recent stimulus history and remote species history are folded onto the dynamic computational state of the system. Observation and decision-making occur simultaneously. It seems, at least, that continuous temporal pattern recognition can only work this way.

I have reviewed various attempts to deal with temporally structured information. We have found that there are only a small number of basic models. The casual assumption of a time window in which measurements are made is not plausible. So simply to state the information that is required by listeners to perceive human language as they do forces more explicit formulation of the kind of system that could extract that information. This task poses fundamental difficulties for perceptual and cognitive models. There needs to be further direct investigation of the dynamic aspects of nervous system behavior. Recurrent network models can play a critical role in research on these issues because they allow systematic comparison of specific hypotheses about the processing of temporal patterns.

References

[Anderson and Port, 1990] Anderson, S. and Port, R. (1990). Experiments with dynamic networks for auditory pattern recognition. Technical Report to appear, Indiana University.

[Bell, 1988] Bell, T. (1988). Sequential processing using attractor transitions. In *Proceedings of the 1988 Connectionist Summer School*, pages 93–102. Morgan-Kauffmann, San Mateo, California.

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

[Clements, 1985] Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, 2:225–252.

[Dorman et al., 1979] 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, 1988] 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 and Zipser, 1988] Elman, J. and Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, 83:1615–26.

[Elman and McClelland, 1986a] Elman, J. L. and McClelland, J. L. (1986a). Exploiting the lawful variability in the speech wave. In Perkell, J. S. and Klatt, D. H., editors, *Invariance and Variability in the Speech Processes*, chapter 17, pages 360–381. Lawrence Erlbaum Assoc., Hillsdale, NJ.

[Elman and McClelland, 1986b] Elman, J. L. and McClelland, J. L. (1986b). 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.

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

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

[H. B. Barlow, 1965] 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 and Stevens, 1980] Halle, M. and Stevens, K. N. (1980). A note on laryngeal features. *Quarterly Progress Report, Research Lab of Electronics, MIT*, 101:198–213.

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

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

[Hawkins and Presson, 1986] Hawkins, H. and Presson, J. (1986). Auditory information processing. In Boff, K. L., Kauffman, L., and Thomas, J., editors, *Handbook of Perception and Performance*, chapter 26, pages 26.1–26.64. Wiley and Sons.

[Hinton, 1987] Hinton, G. (1987). Connectionist learning procedures. Technical Report CMU-CS-87-115, Carnegie Mellon University. To appear in the Journal of Artificial Intelligence.

[Hinton, 1988] Hinton, G. (1988). Representing part-whole hierarchies in connectionist networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, New Jersey. Erlbaum. To appear.

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

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

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

[Kewley-Port, 1983] 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 and Watson, 1989] 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, 1976] 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, 1986] 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.

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

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

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

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

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

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

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

[Port, 1986] 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 and Anderson, 1989] 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 280–288, Hillsdale, NJ. L. Erlbaum Assoc.

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

[Port and Dalby, 1982] 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 et al., 1988] Port, R., Reilly, W., and Maki, D. (1988). Use of syllable-scale timing to discriminate words. *Journal of the Acoustical Society of America*, 83:265–273.

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

[Port and Rotunno, 1979] 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.

[Sankoff and Kruskal, 1983] 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 and Rosenberg, 1986] Sejnowski, T. and Rosenberg, C. (1986). Nettalk: A paralle network that learns to read aloud. Technical Report JHU/EECS-86/01, The Johns Hopkins University Electrical Engineering and Computer Science.

[Smolensky, 1986] Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In Feldman, J. A., Hayes, P. J., and Rumelhart, D. E., editors, *Parallel Distributed Processing*, chapter 22, pages 390–431. The MIT Press, Cambridge, MA.

[Smythe, 1987] 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, 1988] Smythe, E. J. (1988). Temporal computation in connectionist models. Technical Report 251, Indiana University, Computer Science Department, Indiana University, Bloomington, Indiana.

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

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

[Stevens and Blumstein, 1978] 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 and Blumstein, 1981] 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, 1951] Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S., editor, *Handbook of Experimental Psychology*, pages 1–49. Wiley, New York.

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

[Tank and Hopfield, 1987] 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, 1985] 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.

[Victor R. Lesser and Reddy, 1975] Victor R. Lesser, R. D. F. L. E. and Reddy, D. (1975). Organization of the Hearsay-II speech understanding system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23:11–23.

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

[Waibel et al., 1988] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1988). Phoneme recognition: Neural networks vs. hidden Markov models. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 107–110. IEEE.

[Watrous, 1990] Watrous, R. (1990). Modular networks for speech recognition. *Journal of the Acoustical Society of America*, page in press.

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

[Watson and Kelly, 1981] Watson, C. S. and Kelly, W. J. (1981). The role of stimulus uncertainty in the discrimination of auditory patterns. In Getty, D. J. and Howard, J. H., editors, *Auditory and Visual Pattern Recognition*, chapter 3, pages 37–59. L. Erlbaum Assoc., Hillsdale, NJ.

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

[Williams and Zipser, 1988] Williams, R. and Zipser, D. (1988). A learning algorithm for continually running fully recurrent neural networks. Technical Report 8805, ICS, UCSD, La Jolla, CA.