TECHNICAL REPORT NO. 258

Dynamic Speech Categorization with Recurrent Networks

by

Sven Anderson, John Merrill and Robert Port

August 1988

COMPUTER SCIENCE DEPARTMENT
INDIANA UNIVERSITY
Bloomington, Indiana 47405-4101

Dynamic speech categorization with recurrent networks

Svën Anderson, John Merrill*, and Robert Port
Phonetics Laboratory
Department of Linguistics and Department of Computer Science
Indiana University
Bloomington, Indiana 47405

Abstract

Topologies and learning algorithms interact in constraining the performance of networks that attempt to deal with time-varying signals like speech. We trained several connectionist networks to classify the English syllables *ba, da, ga, pa, ta, ka.* Tokens of the six syllables were collected from 10 male and 10 female speakers. Using a speech preprocessor, perceptually scaled spectra and zero-crossing rate were computed every 5 milliseconds and used as an input stream. A subclass of recurrent networks inspired by Jordan (1986) were trained by second order backpropagation to categorize the stimuli. Trained on 10 speakers, a sequential network correctly categorized 87% of the syllables from the other speakers. Moreover, relevant speech cues are extracted rapidly by the network as processing proceeds. We conclude that recurrent networks are well suited to an inherently temporal domain.

1 Speech Categorization

Modern phonetics has taken one of its primary tasks to be that of discovering the acoustic cues that are to be identified with various abstract categories such as allophones, phonemes, syllables, words, and syntactic boundaries. The feature systems of Jakobson, Fant, and Halle (1951), Chomsky and Halle (1968), and Sagey (1986) are intended to specify the primary cues, both acoustic and articulatory, necessary to distinguish the sounds of English. Research in acoustic phonetics, however, suggests that static matrices of acoustic-articulatory features do not adequately specify speech sounds [Liberman et al., 1967]. Instead, our perception of distinct categorically valued phones results from a perceptual process that employs many independently obtained temporally distributed features. The temporal dimension of speech is important not only for its prosodic characteristics (e.g., stress) but also for its segmental description.[1]

Speech is produced by a slowly moving vocal tract. Consequently, speech cues are distributed over time in a complex manner. The cognitive system in humans responsible for phonetic perception must assimilate large numbers of weak cues that are distributed over time in order to obtain a categorical percept. For example, the words 'shop' and 'chop' are perceptually distinct when said in isolation. Dorman et al. (1979) have shown that the word 'shop' in the utterance 'please say shop' is perceived as 'chop' if more than 50 milliseconds (msec.) of silence is introduced between the words 'say' and 'shop'. The vowel in 'say' and following silence, both temporally distant from the frication, constrain higher level categorical perception. The duration of silence is also critical to perception of 'chop' versus 'shop'.

It is this aspect of perception, integration of information distributed over time, that is not adequately handled by feed-forward networks that lack other means to handle time. In the experiments described here we chose the task of identifying stop-vowel syllables differing in

*Center for Adaptive Systems, Department of Mathematics, 111 Cummington Street, Boston University, Boston, Massachusetts 02215.

[1] Klatt (1976) provides an excellent review of segmental timing.

place of articulation and voicing. This problem is difficult because acoustic information specifying place of articulation is available early in the syllable [Stevens and Blumstein, 1978] whereas voicing information depends on acoustic differences much later in the syllable [Lisker and Abramson, 1964].

2 Speech Categorization By Networks

During the past several years connectionist networks have been applied with some success to the problems involved in speech perception. Neural networks excel at simultaneously combining large numbers of cues to detect objects that are constant at a more abstract level. The parallel processing of acoustic cues permits each of them to act as a weak or 'unreliable' knowledge source and to interact with other cues. The value of this general approach is attested by the success of speech recognition systems that use hidden Markov models [Lee and Hon, 1987]. Like connectionist networks, stochastic models can handle the large amount of variability in acoustic cues [Waibel et al., 1988].

2.1 Feed-forward Networks

Elman and Zipser (1988) have examined feed-forward architectures that perform simple speech recognition. Their networks consisted of 320 input nodes, 2 to 6 hidden nodes, and as many output nodes as there were distinct consonant-vowel sequences to be distinguished (usually 9). Elman and Zipser show that this sort of network can be trained to successfully categorize stop-vowel syllables despite variation across hundreds of tokens taken from a single speaker. Feed-forward networks capture temporal information by converting it into a spatial representation. They use temporal windows of speech since without fixing the size of the input representation, they cannot fix the topology of a feed-forward network.

Every digital speech processor must subdivide a continuous waveform into discrete values on some time scale. An important distintion exists between *time slices* and *time windows*. Within the domain of speech, these terms are probably best defined with respect to the rate at which articulatory gestures are made. A 5 millisecond time *slice* of speech will contain only part of even rapid vocal tract gestures (like the intervocalic flap in 'butter'), whereas a single *window* of 30 msec. or more can specify enough information to accurately cue place of articulation in stops [Kewley-Port et al., 1983]. In short, single time slices do not convey enough acoustic information to be phonetically useful—time windows do. Feedforward networks make use of a time window of duration sufficient to capture relevant cues in the longest token [Klatt, 1986]. The duration of the window of speech selected is arbitrary, but *must* span the time segment in which the necessary cues reside. Thus, for most speech tokens the window selected has redundant nodes which remain unactivated by the input for stimuli that are shorter than the maximum. Although node redundancy is not in itself a problem, that redundancy exists merely to fix network topology.

Another problem concerns the representation of temporal dependencies, such as those that hold across several segments [Port et al., 1987; Anderson and Port, 1988]. Elman and Zipser (1988) also describe a network that performs an identity mapping of 9 syllables. They note that as input patterns are shifted through the input layer, the recognition error rate is minimized whenever the utterance boundaries are aligned with the input window boundaries. When the beginning of an utterance is not aligned with the first clique of input nodes, the net error increases very rapidly. The ability of a feed-forward network to exploit temporal dependencies in the data depends critically on the pre-segmentation of input into appropriate windows by a pre-processor. Any cues lying outside the window cannot be used by the network, so the choice of window size can have important implications for the internal representation generated by the network.

Because it has no memory, a feed-forward network is not equipped to represent processing as state transitions over time without mechanisms such as delay links [Waibel et al., 1988; Tank and Hopfield, 1987]. Instead, the temporal dimension must be treated as an additional aspect of input representations. Recoding time as space cannot be the mere conversion of one dimension to another; it is inevitably the transformation of system dynamics into actual connections among nodes. The model suggested by the architecture of recurrent networks is one in which the state of a network at any point in time depends on a complex aggregate of previous states. In a feed-forward network the hierarchical structure of these states is limited by the number of layers it has, and therefore the representation of system dynamics is necessarily limited.

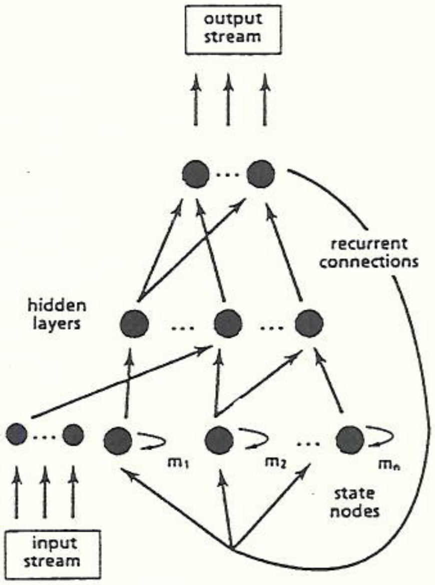

Figure 1: Sequential network for dynamic categorization.

3 Sequential Networks

Recurrent networks offer a partial solution to these problems. In these networks temporal dependencies are not represented by means of network topology, but exist implicitly in the operation of the network. The network used in the experiments to be described is an extension and reinterpretation of the general class of network used by Jordan (1986) to model sequential behavior. On Jordan's view, one can model serial action using a feed-forward network with limited feedback loops. We will refer to this type of network as a sequential network. The input units in this network receive what Jordan calls a *plan*, a constant vector that triggers a certain pattern of sequential behavior in the network. In addition to a layer of hidden nodes, the network contains a group of nodes called *state* nodes that receive input from themselves and the output nodes. State nodes initially have no activation. The weights between output nodes and state nodes have constant value 1. The activation level of the state nodes at an arbitrary time t can be calculated as follows:

$$S_1 = S_0 + O_0$$

and by induction,

$$S_t = S_0 + \sum_{t=1}^{T} \mu^{t-1} O_{T-t}$$

where S is the activation of a state node (index supressed) and O the output activation of a node connected to it. The recurrent connections linking state nodes to themselves act as decay terms, allowing the network to maintain distributed memory of past activation levels.

During execution, Jordan's network operates as follows: the plan and state node activations are applied to the input nodes and all activations are forward propagated until the output nodes are active. The output nodes then directly activate, in turn, the state nodes. Each output vector corresponds to an element of the entire output sequence. Thus, the network can be trained to output various sequences depending on the initial vector of plan activations. As Jordan notes, the nodes of a feedforward network are only activated by nodes in previous layers. Therefore, a feedforward network with constant input can produce only a single response; it cannot generate sequences. If the plan is held constant, recurrent connections are required in order to produce sequential action.

In the networks examined in this paper, the input to the layer of input nodes enters one time slice at a time over many cycles of the network, as shown in Figure 1. The speech signal is inherently dynamic, so a network that accepts a stream of speech requires recurrent connections in order to maintain a representation of previous cues. Each output node of our system communicates with a clique of state nodes, each member of which has a different weight associated with the recurrent connection to itself. The state nodes combine these decaying activations with the ongoing processing at the output layer to retain useful portions of the past states of the network. During training the first hidden layer of a network learns to maximize the useful information it extracts from the state nodes. The network cycles for a number of times specified from outside the system, although in the experiments reported here the network only cycled while the data stream was input (here 25 time cycles).

4 Methods

4.1 Speech Materials

The syllables we chose for this categorization task were single repetitions of the six syllables [ba, da, ga, pa, ta, ka]. An important cue for point of articulation of the stop consonants is the direction in which the second and third formants move toward the relatively steady state of the following vowel [Delattre et al., 1955]. Spectrum shape following release (the point in time at which the vocal tract opens) also contributes to the determination of place of articulation [Kewley-Port, 1983; Stevens and Blumstein, 1978]. Stop voicing is strongly correlated with voice-onset-time, the time from the stop release until glottal excitation [Lisker and Abramson, 1964]. The duration of high zero-crossing rate correlates with the voiceless feature in these consonants, since high zero-crossing rate indicates a random waveform. The stop consonants provide a well-understood data set on which to test the ability of a network to integrate complex information distributed throughout time and perform simple categorization on the basis of that information. Using stimuli from many speakers complicates the task considerably, since each speaker's voice has different spectral and temporal characteristics. The ability of a system to perform cross-speaker categorization is a crucial test of how well a network has grasped the necessary generalizations that make speech possible for humans.

4.2 Subjects

Twenty native speakers of English read 5 numbered lists, each list containing the 18 possible combinations of the consonants [b,d,g,p,t,k] and vowels [a,i,u]. In order to facilitate reading the syllables, each list maintained a specific vowel order and the three combinations with each consonant occupied a single line. Both vowel and consonant order were randomized across the 5 lists, and the lists themselves were shuffled before they were given to the subjects. Subjects read the lists at a relaxed tempo and were recorded in a moderately quiet environment. All syllables having the vowel [a] were extracted from the productions and digitized at 16 kHz through a 6.4 kHz low-pass filter. The syllables were then edited to remove all but the portion of the syllables from 20 msec. before stop release to 100 msec. after stop release. This was done to speed training, since approximately three-fourths of a syllable is composed of steady-state vowel. This part of a syllable is unnecessary for the correct identification of stops by humans.

Before using the stimuli, we performed a simple perceptual study to eliminate those that were ambiguous to listeners. The onsets and offsets of the edited stimuli were digitally ramped over 10 msec. to remove abrupt edges. A single randomized list of the 600 stimuli was prepared and presented binaurally to 13 unpaid listeners over headphones in a forced 6-way categorization task. Listeners were told that they would hear one of the 6 syllables [ba, da ,ga, pa, ta, ka] produced by various people and were asked to write their choice on an answer sheet.[2] Listeners correctly categorized 97% of the productions. A total of 24 stimuli were incorrectly categorized by two or more listeners. These tokens were excluded from the data set used in studying networks. On this revised data set, humans

[2] Four of the subjects were familiar with how the stimuli had been prepared.

correctly categorized 99% of the stimuli.

4.3 Preprocessing

Input tokens to the network were preprocessed in a manner consistent with gross characteristics of the peripheral auditory system. From the edited waveforms, 256 point pitch-synchronous Fourier transforms were computed. The transforms were interpolated onto 5 msec. frames, passed through 35 1-Bark filters placed at half bark intervals [Sekey and Hanson, 1983], and linearly compressed onto the interval [0,1]. The normalization was based on the standard deviation of the entire data set to ensure that more than 95% of all extrema fell between 0 and 1 without compromising the dynamic range of the input data. In addition, zero-crossing rate was generated and normalized in the same manner. Thus, for every syllable the network received a stream of 25 5-msec. spectral slices, each of which consisted of 35 frequency-intensity values and one zero-crossing-intensity value. Due to problems extracting pitch for one of the female speakers, 15 stimuli could not be processed. The total number of syllables used was thus reduced to 561. Stimuli from 10 of the speakers (5 male, 5 female) were randomized and used for training; the other half of the data were used for testing networks.

5 Results and Discussion

5.1 Overall Performance

The output of the networks was specified by Boolean values. In one set of simulations, the activation of each node represented exactly one of the 6 possible syllables. The output layer of the second set of simulations had just 4 nodes: one node represented voicing, and the other three independently represented the three places of articulation. In order to foster learning at every cycle of network processing, target output activations for the correct syllable rose linearly with each 5 msec. time slice from the unbiased value of 0.5 to a final value of 1.0 (following a moving fixed-point technique used by Watrous and Shastri (1987)). Simultaneously, the target activation of other output nodes fell linearly from 0.5 to 0. All networks were trained using second order back-propagation [Parker, 1987]. The differential equations that arose were solved using the trapezoidal rule. Networks were simulated on a VAX 8800.

Networks of many different sizes were tested. As training proceeded, the 3 learning coefficients were lowered. Training ceased when the sum-squared error of the training set no longer decreased by at least one part in 1,000 over 50,000 presentations of the training data. Total training time was usually between 400,000 and 2 million presentations. Most networks learned the training set to a total sum-squared error (SSE) of less than 1000, where all but about 5% to 15% of the training stimuli were correctly categorized.[3] The SSE for the testing set were similar, having percentage correct from 72% to 87%. The best performance for any network was 87% for a network having 4 layers of 66, 66, 12, and 6 nodes. Table 1 summarizes some of the networks that have been tested. Networks with 4 layers appear to have the ability to build an internal representation of the spectrum that leads to better generalization. This would allow normalization for vocal tract variation. It is interesting to note that no 'overlearning' of the training set was observed in any of the simulations. The best solutions to the training stimuli were also optimal for the set of testing data.

5.2 Error Analysis

An examination of errors made by the network reveals a pattern of confusions that mirrors both the known phonetic differences among the stop consonants and the errors human listeners make. Table 2 is a confusion matrix for network 7. Voiced and voiceless stops are rarely confused. When voicing confusions occur, they primarily

[3] For testing data SSE is based on the difference between target and actual activations for every time slice. Throughout this paper we use a winner take all criterion at the final time slice. In the case of the networks having a node representing voicing, values above .5 indicate voicing, and below or equal to .5 indicate voicelessness. Using greatest average activation to interpret network response yields nearly identical results.

Net	# of Layers	Nodes per Layer	Train SSE	Test SSE	% Correct
1	4	66,66,12,6	384.5	594.1	86.9
2	3	72,12,4	463.7	533.8	83.6
3	4	72,72,12,4	818.8	528.2	82.9
4	3	66,10,6	607.5	815.3	81.8
5	3	66,20,6	551.1	858.5	81.4
6	3	66,12,6	616.5	1036.7	78.1
7	3	66,6,6	810.0	988.1	76.6

Table 1: Best scores for 3 and 4-layer sequential networks. Net = network reference number; SSE = sum-squared error.

Network Response

	ba	da	ga	pa	ta	ka
ba	96	0	0	2	0	2
da	9	50	30	0	9	2
ga	9	23	56	0	0	12
pa	2	0	0	94	2	2
ta	0	0	0	11	80	9
ka	0	0	0	13	7	80

Table 2: Syllable confusion matrix for 274 testing stimuli in percent correct. Intended syllables are listed at left.

occur between stops having the same point of articulation. It is well known that because of similarity in production, homorganic stops share acoustic properties (e.g., formant transitions). The greatest source of confusions is obviously between [da] and [ga]. Similar basic results have been observed for every network tested.

Confusions were also analyzed with respect to the behavior of output activations over time. We selected only those responses from the testing set produced by network 7 that were correctly categorized (that is, the diagonal cells of Table 2). At each point in time the network produced an output vector of length 6. For each stimulus, these output vectors were concatenated to yield a description of the network's evolution during the 120 msec. stimulus duration. These 210 vectors were then hierarchically clustered using a Euclidean distance metric and the complete-link method. Figure 2 shows the results of this cluster analysis. One can see that network response to voiceless and voiced sounds are most distinct, whereas [ga] and [da] are most similar. These results show that the relative confusability of the stops is not restricted to the last cycle (on which the confusion matrix in Table 2 is based) but is manifested over a significant part of the stimulus duration.

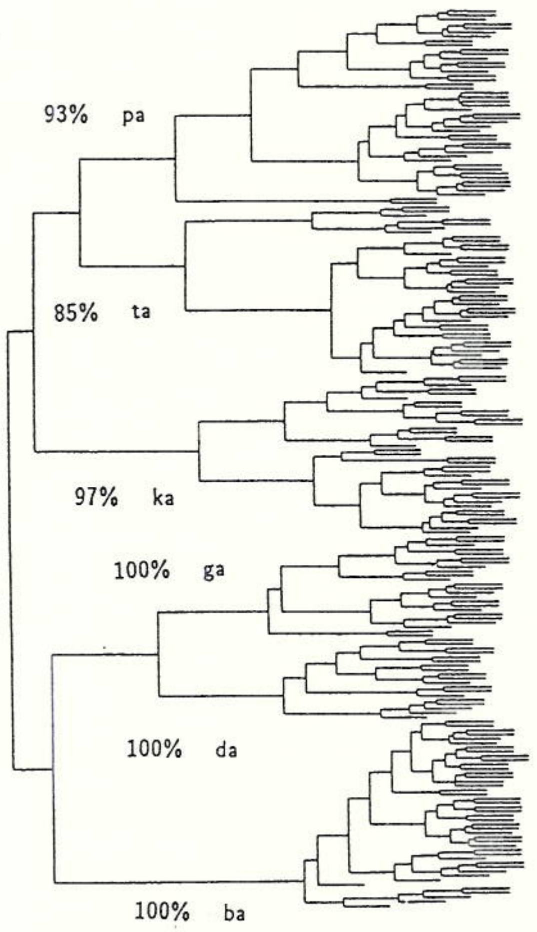

Figure 2: Cluster analysis of 210 testing tokens from 0-120 msec. Branches are labelled with a percentage breakdown of the syllable types they dominate.

This cluster pattern is remarkably similar to the confusion errors made by humans on the same set of syllables masked by varying amounts of white noise [Miller and Nicely, 1955].[4] Shep-

[4] Miller and Nicely examined a much larger stimu-

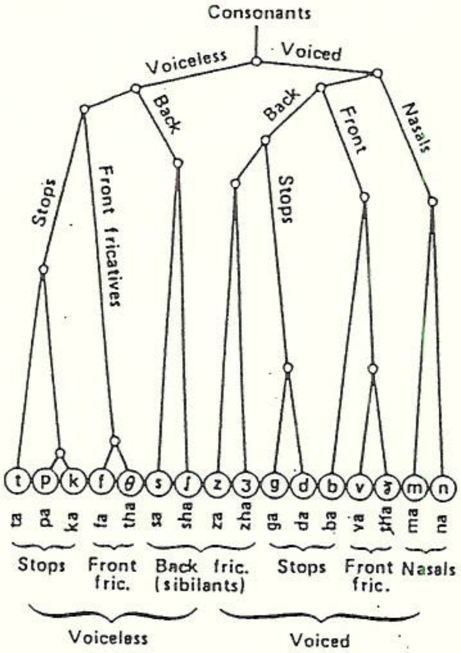

Figure 3: Cluster analysis of the Miller and Nicely data that was not differentially filtered. Reproduced from Shepard (1980).

ard (1980) applied cluster analysis to the Miller and Nicley results. The resulting displays are reproduced in Figure 3. The syllables [da] and [ga] are most confusable for humans and sequential networks alike, while the voicing distinction is salient for both.

In sequential networks, the process of categorization takes place over time. The similarities among output activations over time reveal the emergence of categories as the network cycles. Such an analysis was attempted by analyzing early and late outputs from the total 120 msec. duration of each network response. Figure 4 and Figure 5 show two cluster diagrams based on different segments of network cycles. In the first diagram the vectors begin at the point of release (20 msec.) and continue for 25 msec. Branches that dominate clusters of similar syllables are labelled with the majority category they dom-

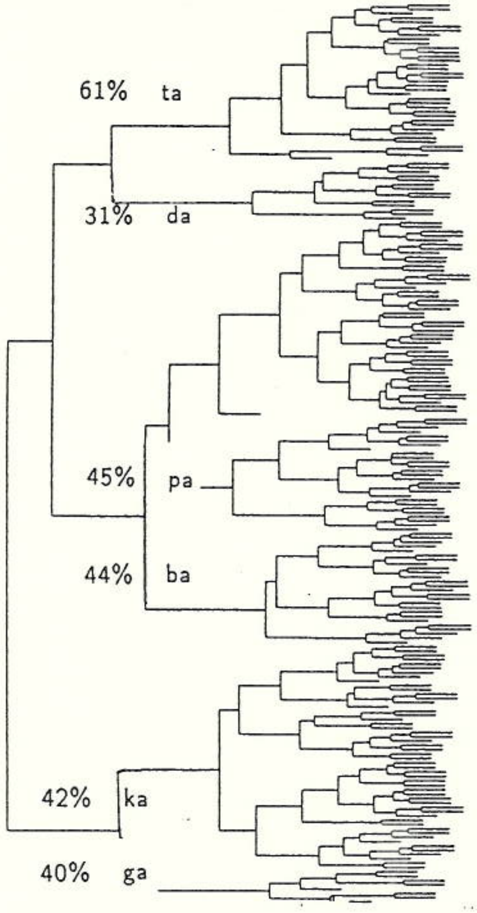

Figure 4: Hierarchical cluster analyses of 20-45 msec. for test stimuli that were correctly categorized.

lus set that included about three quarters of all English consonants.

inate. During this processing interval, the network has begun to differentiate the stops according to point of articulation (labial, alveolar, or palatal). This behavior is consistent with studies demonstrating that humans can use the the first 10-60 msec. of the waveform after release to determine place of articulation in stop consonants [Kewley-Port et al., 1983; Stevens and Blumstein, 1978]. During this interval, voicing is still not determined by the network, because it depends on how soon periodicity appears in the syllable. A comparison of the later portion of the output (80-100 msec. after release) reveals that a voicing decision has emerged. We surmise that sequential networks are capable of rapid extraction of speech cues from an input stream. As we shall see, this is apparently essential to their solution of the categorization task we chose.

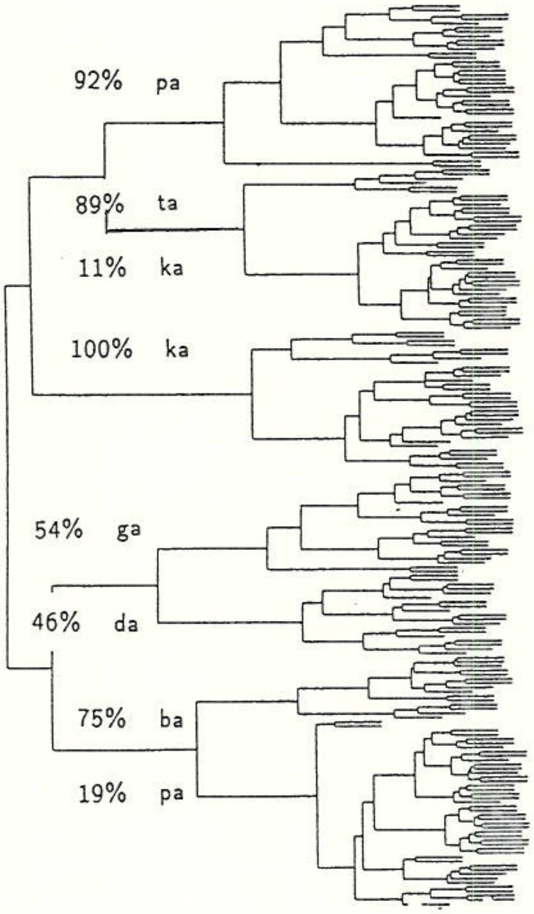

Figure 5: Hierarchical cluster analyses of 80-105 msec. for test stimuli that were correctly categorized.

5.3 Learning in Sequential Networks

Context sensitivity is extremely important in many speech tasks. Beyond merely recognizing appropriate cues, a network that performs in synchrony with stimulus input must allocate a number of its processing elements (state nodes in sequential networks) to serve as memory. If, for example, the network fails to encode early cues bearing on place of articulation, then that information will be lost to the final decision process. In fact, the interpretation of the place of articulation cue has a synergistic effect on the network's use of voicing and other cues. The implications for our choice of learning algorithm are important. As we will show supervised learning algorithms that require training at each cycle will tend to purge the network's memory at each cycle, obstructing the creation of appropriate mechanisms for information retention.

The output representation we employed in networks 1, 4, 5, 6, and 7 does not distinguish the 6 syllables in terms of voicing or place of articulation. All 6 output representations are linearly independent. For example, at any given cycle of a network that is being presented a token of [ba], the target value for the [ba] output nodes is no more similar to the target value for [pa] than it is to any of the other four syllable types. Nevertheless, the actual outputs of the networks share important similarities that clearly capture significant generalizations about stop consonants. This is demonstrated by the activation clusters in Figures 3 and 4, which indicate that the *actual* outputs possess meaningful information (e.g., place of articulation) about that portion of the stimulus the network has already processed. The output activations convey more information to the state nodes than is specified by the target function with which they were trained.

This observation accounts, in part, for the failure of certain training techniques. In training sequential networks that accept constant input vectors, the conventional wisdom is that the target vector should be propagated from output to state nodes in order to speed training (Jordan, personal communication). This is because propagating the actual output rather than the desired output may cause error to be propagated from cycle to cycle. However, actual outputs can provide a means for encoding and storing information. In all of the networks described here, the actual output activations, not the target functions, were propagated back to the state nodes. Several attempts (using both back-propagation and second-order back-propagation) were made to train our networks by propagating target activations to the state nodes during training. None of the networks was successfully trained. We believe the reason for this is that simply specifying the category an output must have prevents necessary information about the history of the signal from being propagated. The network cannot learn the target functions and also perform the task successfully. The target function fails to encode necessary contextual information.

The difficulty of learning the linear ramping function can be seen from the slope of the output activation functions. We have not been able to train a network to achieve a sum-squared error below 300 on the training set. Although this might be due to the limited number of computer cycles available to us, it seems likely to be a consequence of the target function and the supervised training algorithm employed. In addition, the networks do not learn to respond with a linearly increasing activation function. Visual inspection (see Figure 6) shows that output transitions are generally steeper than the slope of the linear ramps they were trained with. In order to demonstrate this statistically, we plotted an ideal linear ramp to the observed maximum value of the actual output for every token. We then calculated smooth derivatives of the actual output activation. We found that 41% of the derivatives were less than half of the slope of the ideal ramping function (to the same maximum activation).[5] This implies that actual activation values rose much more rapidly than those of the target function. The target function is not learned perfectly, and final category decisions must be made rapidly, not slowly.[6]

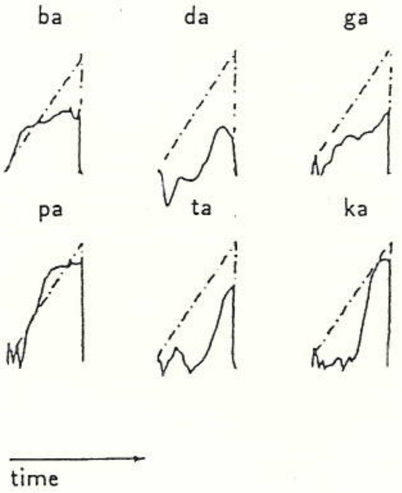

Figure 6: Examples of typical output activations (solid lines) displayed with their target functions (dashed lines) corresponding to 6 stimuli.

[5] Results for the testing data were 44%. Since the network was trained to produce a linear ramping activation for the training data, we emphasize that case here.

[6] Waibel et al. (1988) have suggested a means of avoiding this difficulty for time-delay neural networks. The authors use an error measure that is integrated over the entire time course of network processing. The weights are then readjusted using the average of the weight changes for each cycle. This can be applied to sequential networks by updating weights after several cycles of processing. Preliminary results suggest the usefulness of this update schedule. A network having the save topology as network 1 was trained to 88.7% correct classification of the test data and a training error of 10.84 using update averaged over 5 time slices.

6 Conclusion

We have shown that sequential networks can categorize dynamic data streams such as those involved in auditory perception. Although the performance of sequential networks on novel speech from speakers outside the training corpus does not approach human performance (87% versus 99%), we found that the mapping learned by sequential networks generalizes surprisingly well across speakers. The ability of sequential networks to generalize rests on the extraction and integration of primary cues for voicing and place of articulation. The extraction of relevant cues arises naturally when time is not represented by external mechanisms.[7] Humans are believed to rely on these cues as well. Another important property of real nervous systems is their rapid response to environmental changes. We have shown that sequential networks can extract cues from speech as they arrive in the input stream.

With respect to temporal processing, sequential networks are theoretically superior to feed-forward networks, because they do not require node connections to represent time. Consequently, sequential networks obviate the need for redundant processing nodes, accomplishing similar tasks with fewer processors.

A very serious problem remains with training sequential networks using supervised learning algorithms. Our target function was chosen to be a monotonic function from the unbiased state to the correct output pattern. The crucial problem is with respect to the arbitrary specification of network dynamics. Without knowledge of the appropriate intermediate states of a network, one has no principled method for selecting a target function. In order to learn at each cycle, back-propagation requires immediate feedback (i.e., feedback before nodes are activated by another stimulus). One must specify the output at intermediate stages if learning is to occur there. Specifying the output activations to increase linearly restricts the network to discovering solutions that nearly satisfy those constraints. This severely constrains the space in which optimization occurs. As we have seen, the linear ramping function does not conform to the actual solution discovered by the networks, yet despite these shortcomings, linear ramps foster the discovery of distinctions important to speech.

7 Acknowledgments

This work was supported by NSF grants DCR-8518725 and DCR-8505635. We are grateful to Ron Cole and Robert Brennan for making the CMU speech processing software available to us. We would like to thank Michael Jordan and Geoff Hinton for their commnents on an earlier version of this paper. The assistance of William Mills in collecting the data is gratefully acknowledged.

References

[Anderson and Port, 1988] Anderson, S. and Port, R. (1988). Segmental durations as cues for suprasegmental structure. In preparation.

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

[Delattre et al., 1955] Delattre, P., Liberman, A., and Cooper, F. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27:769–773.

[Dorman et al., 1979] Dorman, F., 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 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.

[Jakobson et al., 1951] Jakobson, R., Fant, G., and Halle, M. (1951). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.* MIT Press, Cambridge, Massachusetts.

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

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

[Kewley-Port et al., 1983] Kewley-Port, D., Pisoni, D., and Studert-Kennedy, M. (1983). Perception of static

[7] Using feed-forward networks with time-delay links, Waibel et al. (1988) have argued that feed-forward networks can show similar tendencies to extract cues such as formant transitions over time.

and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73:1779–93.

[Klatt, 1986] Klatt, D. (1986). Comment on exploiting lawful variability in the speech wave. In Perkell, J. and Klatt, D., editors, *Invariance in Phonetics*. Erlbaum Associates, Hillsdale, New Jersey.

[Lee and Hon, 1987] Lee, K.-F. and Hon, H.-W. (1987). Large-vocabulary speaker-indpependent continuous speech recognition using hmm's. Technical report, Carnegie Mellon University.

[Liberman et al., 1967] Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). Perception and the speech code. *Psychological Review*, 74-76:431–461.

[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.

[Miller and Nicely, 1955] Miller, G. and Nicely, P. (1955). Perceptual confusions among consonants. *Journal of the Acoustical Society of America*, 27:338–352.

[Parker, 1987] Parker, D. (1987). Optimal algorithms for adaptive networks: Second order back propagation, second order direct propagation, and second order hebbian learning. In *ICNN Conference Proceedings*, pages II:593–600. IEEE.

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

[Sagey, 1986] Sagey, E. (1986). *The Representation of Features and Relations in Non-linear Phonology*. PhD thesis, MIT.

[Sekey and Hanson, 1983] Sekey, A. and Hanson, B. (1983). Improved one-bark bandwidth auditory filter. Presented at the 106th meeting of the Acoustical society of America.

[Shepard, 1980] Shepard, R. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398.

[Stevens and Blumstein, 1978] Stevens, K. and Blumstein, S. (1978). Invariant cues for place of articulation in stop consonatns. *Journal of the Acoustical Society of America*, 64:1358–68.

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

[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 *Proceedings from ICASSP*, pages 107–110. IEEE.

[Watrous and Shastri, 1987] Watrous, R. and Shastri, L. (1987). Learning phonetic features using connectionist networks: An experiment in speech recognition. In *ICNN Conference Proceedings*, pages 381–388. IEEE.