Learning Acoustic Features From Speech Data Using Connectionist Networks

Raymond L. Watrous [1]
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
1-215-898-8542
Siemens Research and Technology Laboratories
Princeton, NJ

Lokendra Shastri
Department of Computer and Information Science
University of Pennsylvania

Keywords: connectionist networks, machine learning, speech recognition

March 13, 1987

Abstract

A method for learning phonetic features from speech data using connectionist networks is described. A *temporal flow model* is introduced in which sampled speech data flows through a parallel network from input to output units. The network uses hidden units with recurrent links to capture spectral/temporal characteristics of phonetic features. A supervised learning algorithm is presented which performs gradient descent in weight space using a coarse approximation of the desired output as an target function.

A simple connectionist network with recurrent links was trained on a single instance of the word pair "no" and "go" represented as fine time-scale filterbank channel energies, and successfully learned to discriminate the word pair. The trained network also correctly separated 98% of 25 other tokens of each word by the same speaker. The same experiment for a second speaker resulted in 100% correct discrimination. The discrimination task was performed without segmentation of the input, and *without a direct comparison of the two items.*

A second experiment designed to extended the use of this model to discrimination of voiced stop consonants in various vowel contexts is described. Preliminary results are described in which the network was optimized using a second-order method and learned to correctly classify the voiced stops. The results of these experiments show that connectionist networks can be designed and trained to learn phonetic features from minimal word pairs.

[1]Thanks to Wolfgang Feix, Alex Waibel, Max Mintz and Bruce Ladendorf for helpful discussion.

1 Introduction

Connectionist networks offer significant advantages in addressing problems of machine perception because of their inherently parallel structure, which is well matched to the biological architecture that has served as their paradigm. Their learning capabilities, robust behavior, noise tolerance and graceful degradation are all capabilities which are becoming increasingly well understood and documented [SR86].

The solution of certain perceptual problems requires that the temporal relationships among stimulus characteristics be properly represented. This is especially true in speech recognition, where the relationship between time and frequency is wonderfully complex. One major result from the past thirty years of speech recognition and synthesis research is that it is generally impossible to define speech as a sequence of events with static spectral characteristics. Instead, speech is produced and perceived as a continuous flow of sound, with constantly changing spectral properties. In the production of speech, basic speech units (phonemes) are integrated into a smooth sequence, so that the acoustic boundaries can be very difficult to specify. Moreover, phonemes are often co-produced (coarticulated), so that the phonemes exert a strongly context-dependent interaction. The effect of context is seen in the changes in formant trajectory, duration and energy contours. Thus, the perception of speech depends on the correct analysis of dynamic temporal/spectral relationships.

Many solutions to the problem of speech recognition have been advanced, including signal processing, feature extraction, pattern matching with dynamic non-linear time alignment, linear predictive coding, stochastic modeling, segmentation and labeling, syntactic grammars and expert systems, with explicit rule-based knowledge representation. These approaches share the goal of capturing the regular structure inherent in the speech signal in the presence of tremendous variability. Although these techniques have all succeeded to some extent, a general shortcoming has been that minor irregularities in the input, whether from signal noise, background acoustic noise, or speaker variability, have major negative effects on performance. This lack of robustness has been very frustrating, because it is contrary to our experience of speech communication, in which minor irregularities are easily overcome, if consciously perceived at all.

The connectionist network approach is attractive because it offers a computational model which has inherently robust properties. The networks consist of simple processing elements which integrate their inputs and broadcast the results to the units to which they are connected. Thus, the network response to input is the aggregate response of many interconnected units. It is the mutual interaction of many simple components that is the basis for robustness.

Connectionist networks also provide a fundamentally different language for knowledge representation. This is important in establishing a conceptual framework in which solutions can be conceived and investigated. In addition, the computational speed of parallel networks is a requirement for real-time performance in non-trivial speech systems.

The problem of designing connectionist networks which can learn the dynamic spectral/temporal characteristics of speech has not yet been widely studied. Most work in connectionist networks so far has focussed on the static relationship between input/output pairs, such as associative memories [KL81,Hop82], various encoding, decoding, parity and addition problems [RHW86], and mapping from word spelling to phoneme labels [SR86].

The TRACE model [EM86,ME86] is the first well-developed model for studying speech recognition using connectionist networks. As discussed below, this model represents temporal sequence directly using sets of network units allocated to subsequent time slices. The approach developed in this paper is different in that temporal sequence is represented implicitly.

Learning to associate static input/output pairs can be accomplished with layered connectionist networks with feedforward links alone. But recurrent, or feedback, links are required to

provide the network with state sequence information, in order to capture sequential behavior. [Jor86,Sut85,RHW86].

The experiments reported here were designed to explore the capabilities of parallel networks to learn *dynamic properties of time-varying data.*

We first choose a moderately difficult speech recognition problem to test the extent to which a connectionist network could form an internal representation of the temporal/spectral characteristics which distinguish two similar words. A simple network with recurrent links was trained on a single instance of the word pair "no" and "go", and successfully discriminated 98% of 25 other tokens of each word for the same speaker. The experiment was repeated for a second speaker and resulted in 100% discrimination performance.

This research is being extended to more difficult problems of speech recognition. Preliminary results are reported which show that connectionist networks can be used to successfully discriminate the voiced stop consonants, /b,d,g/, in various vowel contexts.

The results of the preliminary experiments show that connectionist networks can indeed be designed and trained to successfully discriminate similar word pairs by learning acoustic-phonetic features.

2 Experiment I

The experiment selected for this first examination of connectionist networks in speech recognition was the discrimination between the minimal pair "no" and "go". This is a typical speech recognition problem, which is included in a standard database for evaluation of speech recognizers [DS81]. The utterances "no" and "go" share for the major and final portion the voiced phoneme /o/. The "no" utterance is characterized by a lower energy nasal murmur preceding the transition to the back vowel /o/. The "go" is distinguished by a very low energy voicing interval during the lingua-palatal closure, a brief burst as the closure is released, and a voiced transition to the full vowel.

The distinction between "no " and "go", therefore, is concentrated in the brief interval of relatively low energy at the beginning of the word. These differences consist in the relative voicing energy, burst spectrum, and formant value and transition pattern.

2.1 Network Architecture

For this first experiment, a three-layer connectionist network consisting of an input layer, one hidden layer and an output layer was implemented, as shown in Figure 1. The sampled speech data flowed through the network in time sequential order. Thus, the 16 channel energies were applied to 16 input units, from which activation spread toward the output units simultaneously as the input units were updated by subsequent speech samples. This design will be referred to as the *temporal flow model*, or, more simply as the *flow model.*

Other approaches have used an array of input units, and represented the time axis along one index of input unit array [PNH86,EM86,ME86]. In these cases, time is spatialized across units. The temporal flow model was chosen because it does not require 'chunking' of variable length utterances onto a fixed size network, it avoids the problem of temporal symmetry, and the temporal flow model seems to be closer to the biological model of speech processing.

Integration over time of spectral characteristics is accomplished by the recurrent unit links. A positive recurrent link weight will feed back as input some of the unit output to reinforce and integrate the unit response.

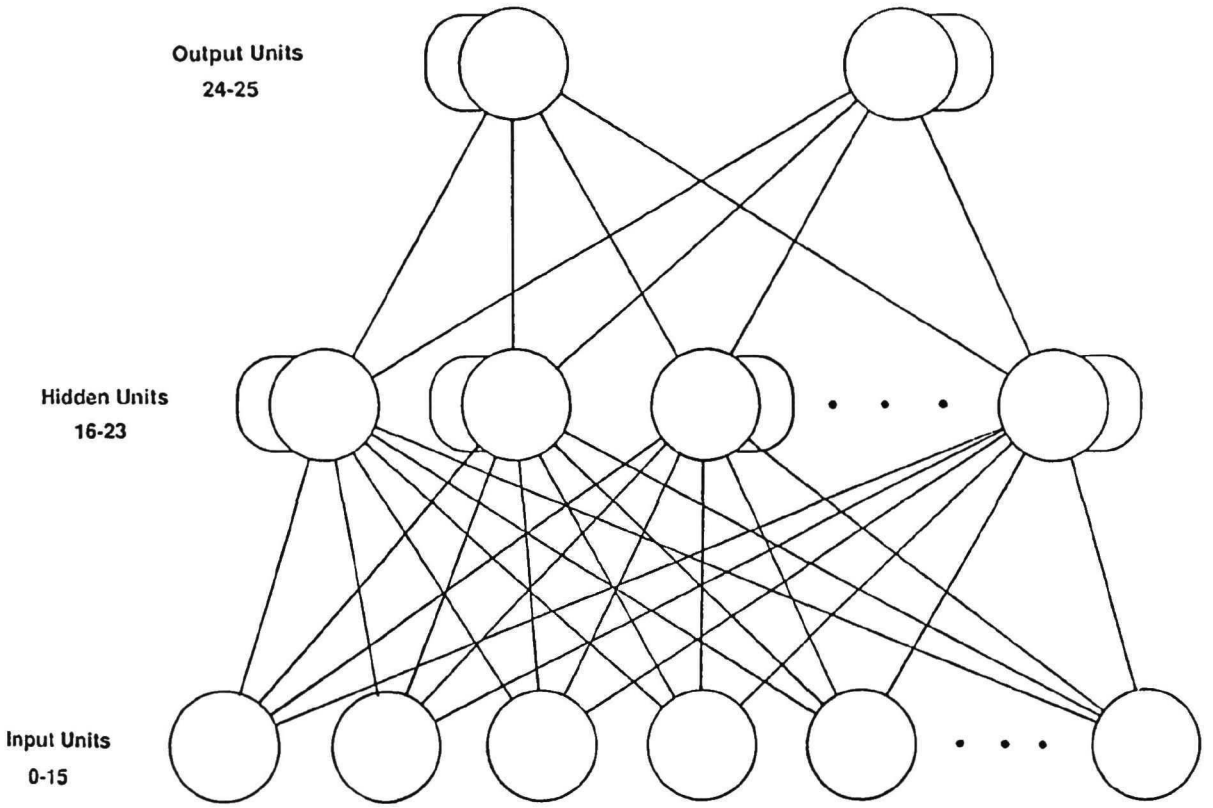

Figure 1: "Temporal Flow Model showing input, hidden and output layers"

2.1.1 Unit Functions

The functions which define the unit behavior were chosen from ones in common use in connectionist networks [SR86,RHW86]. These functions approximate the computational properties of neural cells, and have convenient mathematical properties for the learning algorithm used in this experiment.

The unit output, $o_j(t)$, is given by the sigmoid function:

$$o_j(t) = \frac{1}{1 + e^{-p_j(t)}}$$

where $p_j(t)$, the potential function is given by:

$$p_j(t) = \sum_{i,d} w_{ijd} o_i(t - d)$$

where d is the time delay along the link between units u_i and u_j.

2.2 Back-Propagation Learning Algorithm

For this experiment, an extended form of the back-propagation learning algorithm was chosen to accommodate networks with recurrent links [RHW86]. The derivation of a more general form of the algorithm for variable delay and recurrent links is given in [WS86].

The error-propagation algorithm modifies the unit connection weights in order to minimize the mean squared error between the actual and desired output values. The weight change rule can be written as:

$$\Delta w_{ijd} = \eta \sum_{\tau} \delta_j(t - \tau) o_i(t - \tau - d)$$

where $\delta_j(t - \tau)$ is the error signal at unit j at time $t - \tau$, with respect to the target values at the output units at time t. This error is given by:

$$\delta_j(t - \tau) = \sum_{\alpha,k} w_{jk\alpha} \delta_k(t - \tau + \alpha) \frac{\partial o_j}{\partial p_j}(t - \tau)$$

for $\alpha \leq \tau$.

The error signal for an output unit is defined by the difference between the actual and target values, times the unit function slope at time t:

$$\delta_j(t) = (o_j(t) - targ_j(t)) \frac{\partial o_j}{\partial p_j}(t)$$

The target function for the output units consists of a simple ramp. For the output unit which corresponded to the utterance being trained, the ramp increased from a value of 0.5 to 1.00 over the duration of the utterance. The other unit was correspondingly decreased from 0.5 to 0. This represented the intuition that evidence for or against a particular word accumulates over its duration, and reaches a level of confidence after the utterance is completed.

2.3 Data

The data used for this experiment consisted of speech data for one male (GD) and one female speaker (CP) from the Texas Instruments standard isolated word recognition database [DS81]. The digitized data was played through an A/D converter (Digital Sound Corporation DSC 2000)

into a commercial speech recognition device (Siemens CSE 1200), where it was passed through a 16-channel filter bank, full-wave rectified, log compressed and sampled every 2.5 milliseconds. The filters were low Q bandpass filters, with linear-log spaced center frequencies [Mar70]. Twenty-six repetitions of each word comprise the corpus, for a total of fifty-two utterances (26 "no" and 26 "go") for each speaker. The filter bank response to the training utterances is shown in Figure 2.

2.4 Results

The connectionist network experiments were conducted on a sequential machine using a network simulator, written specifically for this experiment. The experiments were carried out on a VAX 8650 and a SUN 3/165 workstation. The network described previously was initialized with small random link weights and trained on a single pair of no/go utterances for 6000 training iterations. Each speaker's data was used to optimize separate networks.

The results of the optimization are shown in Figure 3 for the first speaker. The value of the squared-error term is neither monotonic decreasing nor a smooth function of the number of optimization iterations. This is thought to be due to the local nature of the weight change algorithm, and the limited extent of back-propagation in time. The network coinciding with the sharp notch in the squared error value was chosen for further study.

2.4.1 Output Unit Response to Training Data

The response of the output units for the network at the selected critical point in the learning process was recorded, and can be seen in Figure 4. The output units respond in equal and opposite ways to the input stimuli; in addition, their time response roughly approximates a ramp. Since the learned response closely fits the training function, the network shows very good discrimination between the single pair of the training set.

The significance of this result should not be overlooked. First, the application locally of global optimization metric provided a successful optimization path to a desired network response pattern. Second, although no segmentation decisions were made, the network was able to form discriminating spectral features independently. Third, the approximations of constant weight value, and restrictions to maximum τ value in the extended back-propagation algorithm did not prevent convergence to a good solution. Fourth, although the shape of the error contour is unknown, it is almost certainly not smooth; consequently, the learning path apparently avoided local minima in arriving at a solution.

2.4.2 Extension to Test Set

In order to test the generality and robustness of the internal representations obtained from the training word pair, and to further investigate the characteristics and behavior of the hidden units, the network of least squared error value was tested on a set of 25 additional pairs of no/go utterances by the same speaker. Using a simple deterministic decision algorithm, the input word could be clearly categorized by the network response. Under these conditions, the trained network successfully discriminated all but one of the test cases (98% accuracy). The results for the second speaker were similar (100% accuracy).

The responses of the hidden units were analyzed for the 50 test utterances as well as the 2 training utterances for each speaker. In nearly every respect, the hidden unit responses of the test utterances were isomorphic to the response to the training data. A single hidden unit provided the discriminatory response. In the single error case, this single unit failed to respond to the input

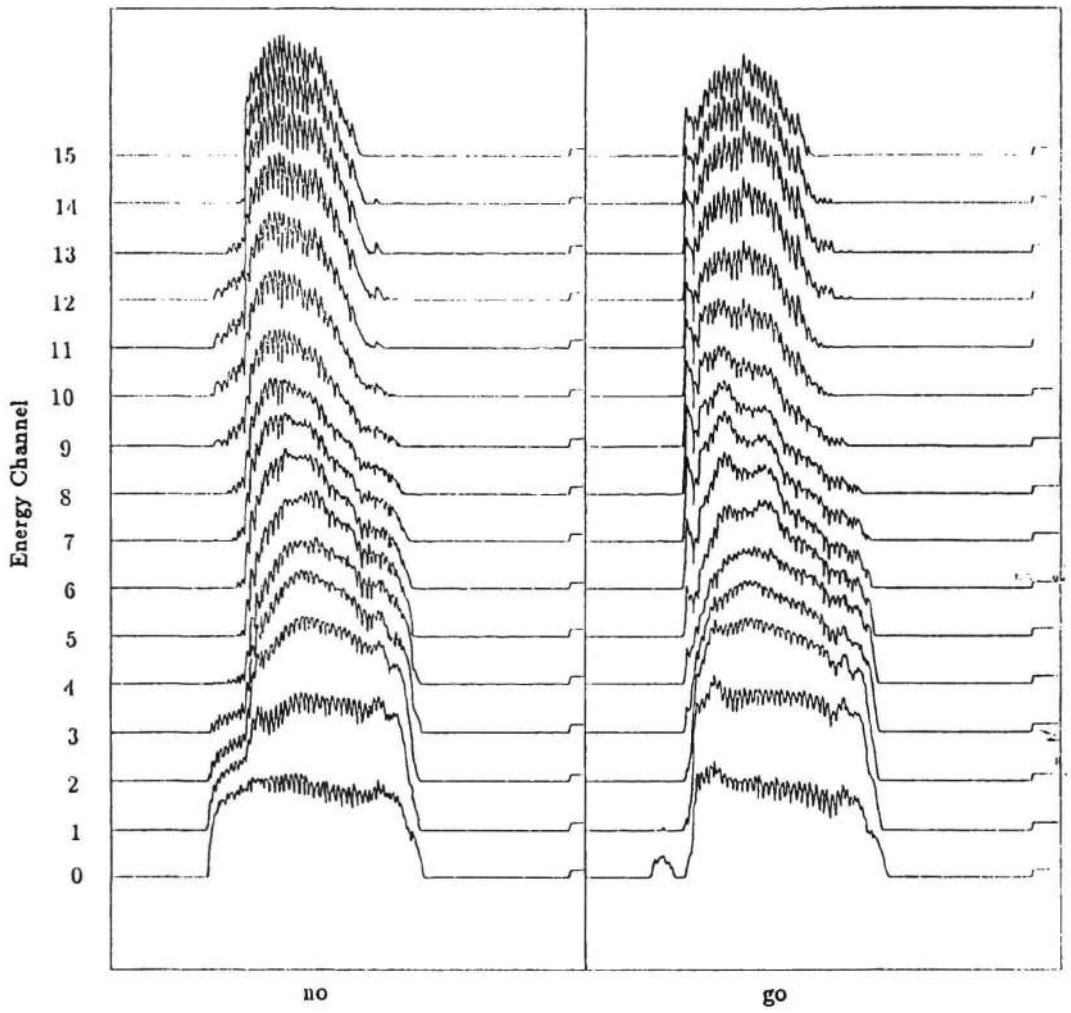

Figure 2: "Channel Energies for no/go pair"

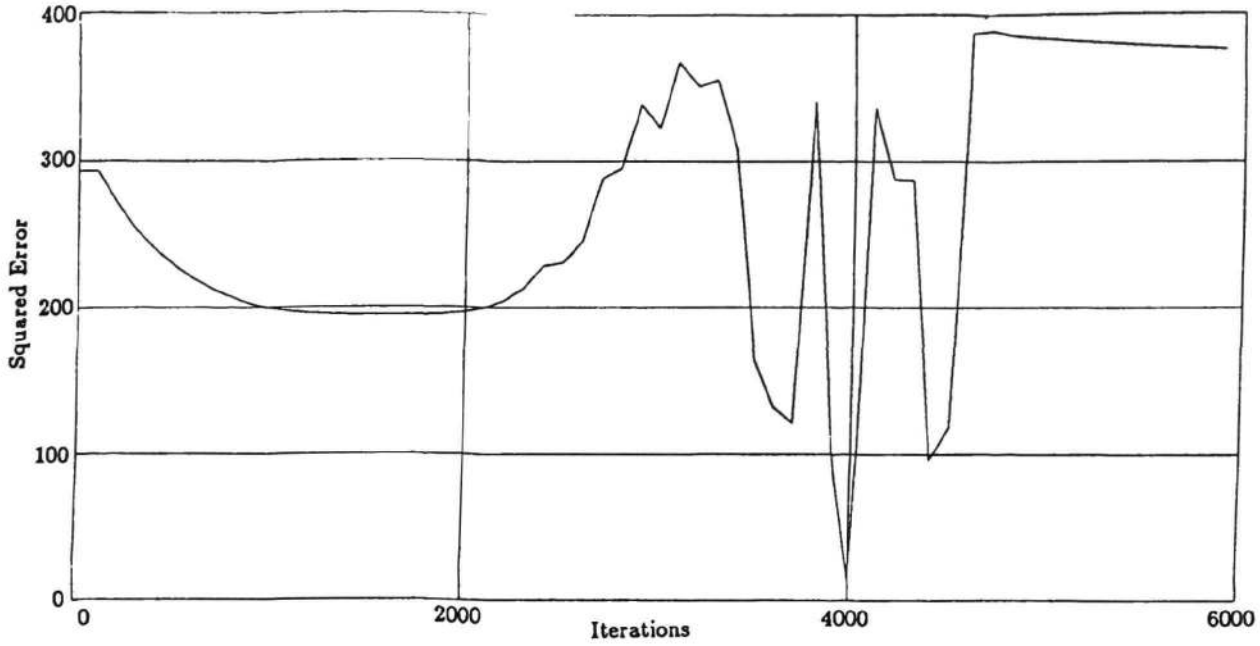

Figure 3: "Squared Error for Training Set vs. Iteration"

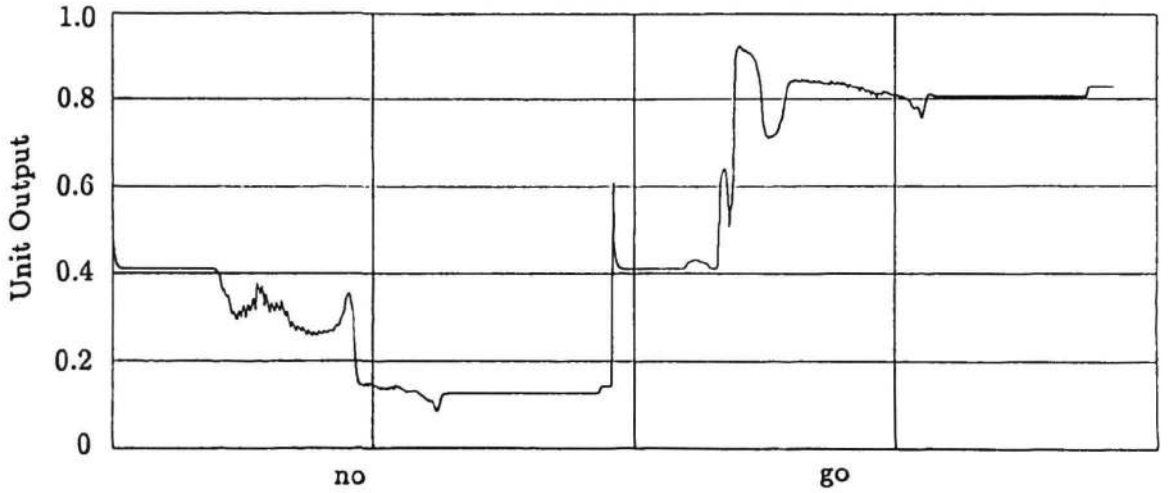

Figure 4: "Output Unit 24 Response to No/Go Pair"

data. The channel energy input data for this utterance is extremely low level, especially in the mid to upper channels.

Based on these encouraging results, a second set of experiments was designed to explore extensions of the use of connectionist networks for more difficult problems in speech recognition.

3 Experiment II

The next set of experiments were designed to learn acoustic-phonetic properties of the phonemes in the category of the voiced stop consonants. The method, however, is completely general and will be extended to other classes of speech sounds. The experimental design is presented, followed by the network architecture and learning algorithm.

3.1 Design of Experiments

The plan of the experiments to learn acoustic properties of phonemes uses a principle of incremental optimization. An initial network is optimized to discriminate the phonemes /b,d,g/ in a particular vowel context, say /i/. When the network has been modified to correctly discriminate the CV words /bi,di,gi/, the training data is expanded to include another vowel context.

The subsequent vowel context for optimization was chosen by two methods. In the first, the subsequent vowel was chosen to be phonetically close to the first. This is designed to test the extent to which the network is able to generalize the consonant discrimination across vowel contexts. The second method was to choose a subsequent vowel phonetically far from the first vowel. This is designed to test the extent to which the network could make context-specific consonant discriminations.

The choice of subsequent vowels was done to increase the likelihood of successful optimization by minimizing the incremental learning required. A series of incremental experiments is defined in this way, by which increasingly dissimilar contexts are added to the network for invariant discrimination of /b,d,g/ and increasing similar contexts are subtracted from the network design to respond selectively to /bi,di,gi/.

An alternative design would involve optimizing the respective networks over the complete data set of positive and negative samples in a single step. This approach has the advantage that the network is forced from the start to attend to the desired goal; a weakness of the incremental approach is that the network could be optimized for one context using an internal representation which is inappropriate for the larger task. In this case, the network would be required to "unlearn" the original solution, which could prove a difficult optimization problem. Nevertheless, the incremental approach was selected for these initial experiments because the progress of optimization could be more closely controlled and evaluated.

It is clear either the subtractive or the additive case would be sufficient for speech recognition. Either context-dependent or context-independent recognition would achieve the goal of consonant discrimination. There is, however, an important theoretical question at issue here of acoustical invariance [Blu86,BS79,BS80,LGB84], which will be explored in a subsequent paper. The results obtained to this point do not warrant a full discussion of the issue of invariance.

3.2 Network Architecture

For these preliminary experiments, a three-layer temporal flow model was implemented, as shown in Figure 1, with a third output unit to accommodate the three voiced consonants.

3.3 Network Learning Algorithm

For these experiments, a second-order optimization algorithm was selected called the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) [Fle80]. This algorithm combines a linear search along a minimizing vector with an approximation of the second-derivative of the objective function f.In this way, knowledge about the structure of the error surface is used to select optimal search directions and achieve much more rapid convergence, especially in the neighborhood of the minima of the objective function. Although such second-order methods do not share the locality property of first-order methods, the BFGS algorithm was employed for the purposes of more rapid optimization using a sequential machine.

The BFGS update formula is given as:

$$H_{(k+1)} = H_k + \left(1 + \frac{\gamma^T H_k \gamma}{\delta^T \gamma}\right)\frac{\delta\delta^T}{\delta^T\gamma} - \left(\frac{\delta\gamma^T H_k + H_k\gamma\delta^T}{\delta^T\gamma}\right)$$

where H is the approximate inverse of $G = \nabla^2 f(\vec{w})$, and:

$$\gamma = \vec{g}_{(k+1)} - \vec{g}_{(k)}$$

$$\delta = \vec{w}_{(k+1)} - \vec{w}_{(k)}$$

The algorithm basically iterates through three steps, as follows:

1. compute the search direction as $\vec{s} = -H\vec{g}$.
2. execute a linear search along $\vec{s}$; that is, minimize $f(\vec{w} + \alpha\vec{s})$ over α, a scalar, $\alpha > 0$.
3. update H according to the BFGS formula above.

The computation of the gradient vector $\vec{g}$ was accomplished by an extended form of the back-propagation learning algorithm for networks with recurrent links as described above.

For these experiments, the initial value of H was chosen to be I. In cases where the linear search failed, H was reset to I, and the search continued.

The linear target function described in the previous experiments was also used for the consonant discrimination experiments.

3.4 Data

The speech data used for the second set of experiments was taken from a small database of isolated consonant-vowel (CV) utterances for a single speaker (RW) consisting of the stop consonants (/p,t,k,b,d,g/) in combination with ten vowels (/i,I,e,ae,a,^ ,o,u,U,3/). Five repetitions of each CV word for a total of three hundred utterances were recorded on a Nakamichi Model 480 tape recorder using the Digital Sound Corporation Model 240 preamplifier. The recorded speech was played into a commercial speech recognition device (Siemens CSE 1200) where it was filtered and sampled as described above. Additional data from the original and other speakers will be collected for further experiments.

The data files were segmented by hand to extract the transition portion of the CV word. The initial segmentation boundary was set at a point of silence at least 50 ms prior to the consonant release and the final segment boundary at the point of maximum vocalic energy, approximately in the center of the vowel nucleus. This segmentation was performed without difficulty and did not involve an attempt to identify the consonant-vowel boundary. The segmentation was done primarily to decrease the computational load on the optimization algorithm. It is certain that sufficient if not complete discriminatory information remained in the segmented data.

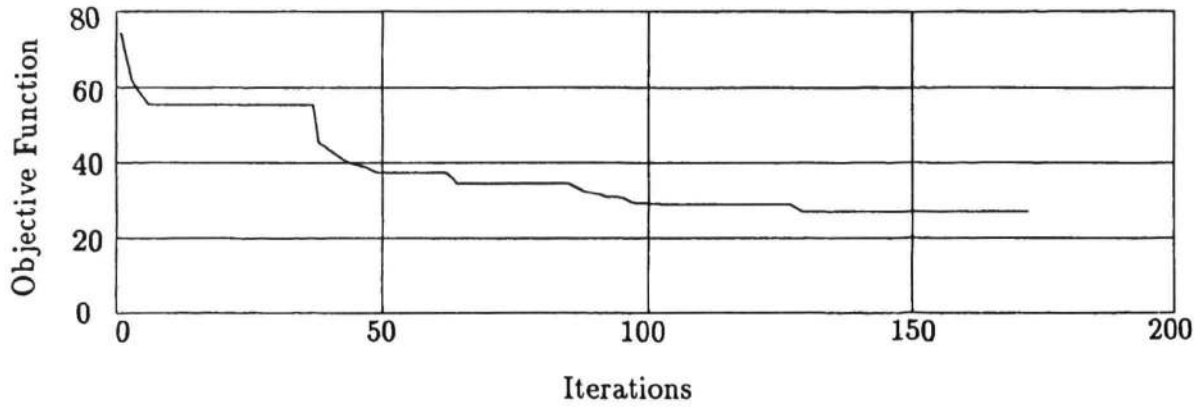

Figure 5: "Objective Function During Optimization"

4 Results

The network described previously was trained on a single set of /bi,di,gi/ using the back-propagation method for networks with recurrent links used in experiment I. The network converged after approximately 3000 iterations.

The resulting network was then trained in two experiments, an additive experiment incorporating /bI,dI,gI/ and a subtractive experiment reducing by /bu,du,gu/. These experiments were conducted with the BFGS algorithm described above. In both cases, the networks converged for proper discrimination in both contexts.

The objective function value during optimization for the subtractive experiment is shown in Figure 5. The response of the output units to the positive training sample for the optimized network can be seen in Figure 6.

The network makes an unambiguous discrimination between the voiced stop consonants in the response of the output units. The unit responses roughly approximate the target function and clearly begin discriminating responses at the initial word boundary.

5 Discussion

The results of the "no/go" experiment have been discussed at length elsewhere [WS86]. In summary, the network formed an internal representation of an acoustic-phonetic feature characteristic of the burst-release of the velar consonant /g/. In addition, the network formed a similar discriminatory mechanism for both speakers. This discriminatory feature was quite robust across repetitions by the same speaker of the same word. Taken together, these two facts strengthen the conclusion that connectionist networks can be used to infer significant acoustic-phonetic features directly from real speech data. This suggests that connectionist learning may be used to uncover acoustical characteristics of the speech waveform in which computational optimality may be found to have perceptual significance.

The results of the consonant discrimination task are preliminary; it is significant, however, that the temporal flow model was correctly optimized to make this discrimination in the context of a single vowel, and that this discrimination was extended additively and subtractively to the nearest corresponding phonetic contexts. Further testing of the invariance hypothesis in the context of

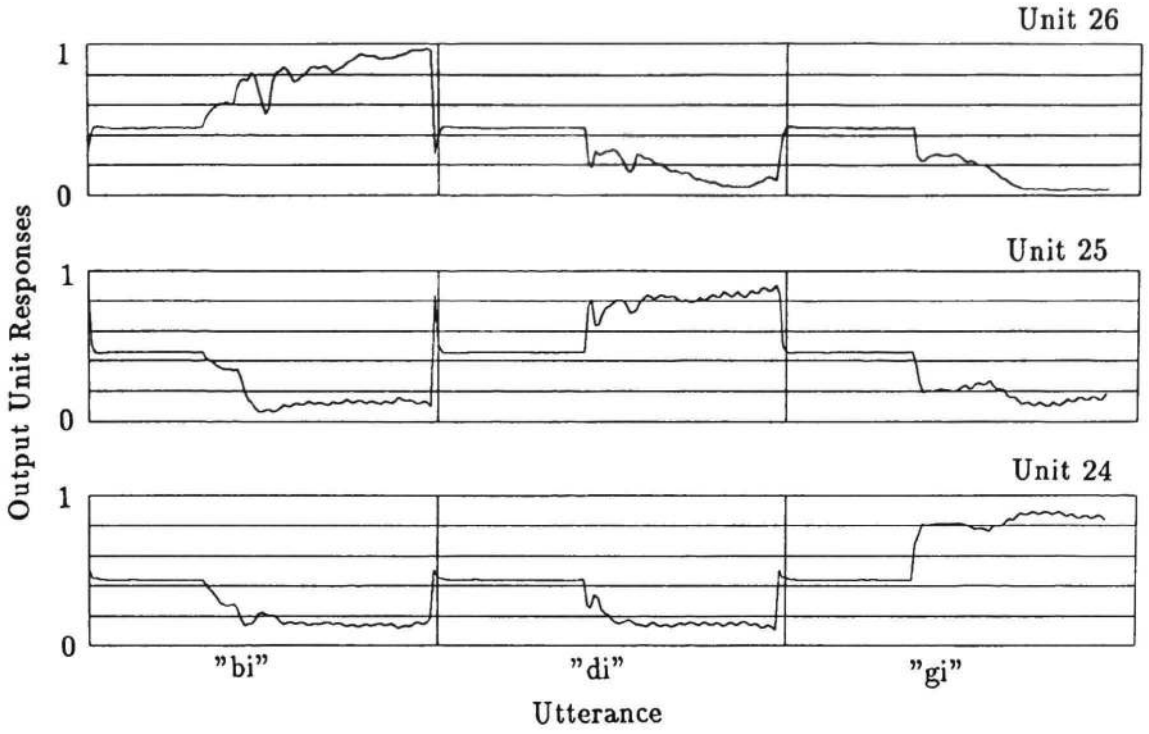

Figure 6: "Output Unit Responses to Training Data"

connectionist networks is currently in progress.

References

[And86] Charles William Anderson. *Learning and Problem Solving with Multilayer Connectionist Systems.* PhD thesis, University of Massachusetts, September 1986.

[Blu86] Sheila E. Blumstein. On acoustic invariance in speech. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*, chapter 9, pages 178–201, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[BS79] Sheila E. Blumstein and Kenneth N. Stevens. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, 66(4):1001–1017, October 1979.

[BS80] Sheila E. Blumstein and Kenneth N. Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 67(2):648–662, February 1980.

[DS81] George R. Doddington and Thomas B. Schalk. Speech recognition: turning theory into practice. *IEEE Spectrum*, 26–32, September 1981.

[EM86] Jeffrey Elman and John McClelland. Exploiting lawful variability in the speech wave. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*, chapter 17, pages 360–380, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Fle80] Roger Fletcher. *Practical Methods of Optimization.* Volume 1 Unconstrained Optimization, John Wiley, New York, 1980.

[Hop82] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the Natural Academy of Sciences USA*, 79:2554–2558, 1982.

[JFH63] R. Jakobson, Gunnar Fant, and Morris Halle. *Preliminaries to Speech Analysis.* MIT Press, Cambridge, MA, 1963.

[Jor86] Michael I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Hillsdale, NJ, 1986.

[KL81] Tuevo Kohonen and Pekka Lehtio. Storage and processing of information in distributed associative memory systems. In G.E. Hinton and J.A. Anderson, editors, *Parallel Models of Associative Memory*, pages 105–143, Lawrence Earlbaum Associates, Hillsdale, N.J., 1981.

[LGB84] Aditi Lahiri, Letitia Gewirth, and Sheila E. Blumstein. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence from a cross-linguistic study. *Journal of the Acoustical Society of America*, 76(2):391–404, 1984.

[Mar70] Thomas B. Martin. *Acoustic Recognition of a Limited Vocabulary in Continuous Speech.* PhD thesis, University of Pennsylvania, 1970.

[ME86] John L. McClelland and Jeffrey L. Elman. Interactive processes in speech perception: the trace model. In J.L.McClelland D.E.Rumelhart and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume II Psychological and Biological Models*, chapter 15, MIT Press, Cambridge, MA, 1986.

[PNH86] David C. Plaut, Steven Nowlan, and Geoffrey Hinton. *Experiments on Learning by Back Propagation.* Technical Report CMU-CS-86-126, Carnegie-Mellon University, 1986.

[RHW86] David E. Rumelhart, Goeffrey Hinton, and Ronald Williams. Learning internal representations by error propagation. In J.L.McClelland D.E.Rumelhart and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume I Foundations*, chapter 8, MIT Press, Cambridge, MA, 1986.

[SB78] Kenneth N. Stevens and Sheila E. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5):1358–1368, 1978.

[SR86] Terrence J. Sejnowski and Charles R. Rosenberg. *NETtalk: A Parallel Network that Learns to Read Aloud.* Technical Report JHU/EECS-86/01, Johns Hopkins University, 1986.

[Sut85] Richard S. Sutton. The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ, 1985.

[WS86] Raymond L. Watrous and Lokendra Shastri. *Learning Phonetic Features Using Connectionist Networks: An Experiment in Speech Recognition.* Technical Report MS-CIS-86-78, University of Pennsylvania, October 1986.