Learning Phonetic Features using Connectionist Networks

Raymond L. Watrous
Siemens Research and Technology Laboratories
Princeton, NJ *

Lokendra Shastri
Department of Computer and Information Science
University of Pennsylvania *

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 evaluation function.

A simple connectionist network with recurrent links was trained on a single instance of the word pair "no" and "go", and successful learned a discriminatory mechanism. The trained network also correctly discriminated 98% of 25 other tokens of each word by the same speaker. A single integrated spectral feature was formed without segmentation of the input, and *without a direct comparison of the two items.*

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 mi.

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. 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. Thus, the perception of speech depends on the correct analysis of dynamic temporal/spectral relationships.

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.

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 [6,4], various encoding, decoding, parity and addition problems [10], and mapping from word spelling to phoneme labels [11].

Learning to associate static input/output pairs can be accomplished with layered connectionist networks with feed-forward links alone. Learning pattern sequences requires network state information, which can be provided by feedback from the network output to the input [6,5,12,10,9]. The idea of learning pattern sequences has been applied to a speech task using Boltzmann machines [9].

The experiments reported here were designed to explore the capabilities of parallel networks to learn *dynamic properties of time-varying data.* We choose a standard 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 network architecture was selected in which the hidden and output units included self-recurrent links. This approach is distinguished from the pattern sequence approach in that the feedback is internal to the network and distributed. Thus, the dynamic response of individual units must be learned in solving the discrimination task.

•Thanks to Alex Waibel for helpful discussion and analysis.
This research was supported in part by DARPA grants N00014-85-K-0018 and NOOO14-85-K-0807, NSF grants DCRr86-07156\ DCR8501482, MCS8219196-CER, MCS8207294, 1 RO1-HL-29985-01, U.S. army grants DAA6-29-84-K-0061, DAAB07-84-K-F077, U.S. Air Force grant 82-NM-299, AI Center grant NSF-MCS-83-05221, U.S. Army Research Office grant ARO-DAA29-84-9-0027, Lord Corporation, RCA and Digital Equipment Corporation.

2 Experiment

The discrimination between the minimal pair "no" and "go* is a typical speech recognition problem, which is included in a standard database for evaluation of speech recognizers [1], 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/. This nasal murmur has a formant structure which is due to the coupled resonances of the closed oral cavity and open nasal cavity. The "go*1 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 't 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 Data

The data used for this experimental work consisted of speech data for a single speaker *(GD)* from Texas Instruments standard isolated word recognition database [1]. The speech data was played 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. Twenty-six repetitions of each word comprise the corpus, for a total of fifty-two utterances (26 "no" and 26 "go"). The filter bank response to the training utterances is shown in Figure 1.

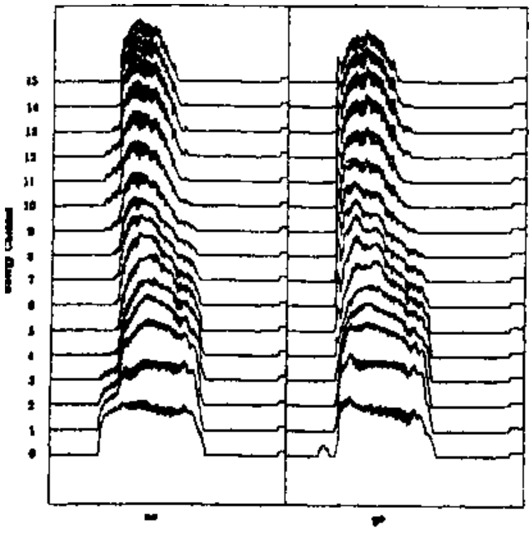

Figure 1: "Channel Energies for no/go pair"

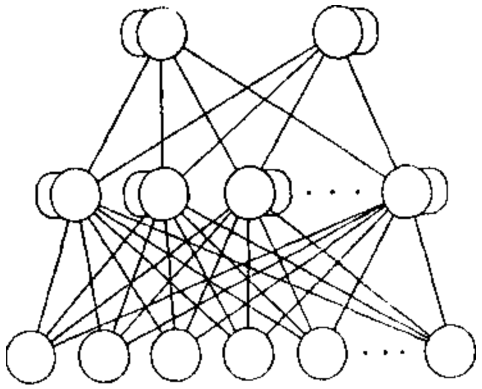

Figure 2: "Network Configuration showing input, hidden and output layers"

2.2 Network Architecture

For this initial experiment, a three-layer connectionist network consisting of an input layer, one hidden layer and an output layer was implemented, as shown in Figure 2. 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 sequential 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 time along one index of the input unit array [8,2,3,7]. In this case, 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 alignment and symmetry, and the temporal flow model seems to be closer to the biological model of speech processing.

2.2.1 Unit Functions

The functions which define the unit behavior were chosen from ones in common use in connectionist networks [11,10]. The unit outputis a nonlinear (sigmoid) function of the unit potential, which is a simple weighted sum of the output values of units connected by afferent links. The weights correspond roughly to the effect of synaptic strengths. The sigmoid function has the desirable properties of a bounded output, non-linear characteristics, and a response threshold. These functions approximate the computational properties of neural cells, and have convenient mathematical properties for the learning algorithm used in this experiment.

2.2.2 Back-Propagation Learning Algorithm

For this experiment, an extended form of the back-propagation learning algorithm was chosen to accommodate networks with recurrent links [10,13].

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:

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

where *6j(t* — r) is the error signal at unit *j* at time *t* - *T,* with respect to the target values at the output units at time *t* [13]. 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 *a* ≧ 1-

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 value of r was limited to a small value to limit the recursive computation. The weight changes were made after each time step. These factors introduced approximations into the computation of the gradient.

2.2.3 Target Function

The target function for the output units used in the no/go discrimination experiment consisted 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.

3 Results

The parallel connectionist network experiments were conducted on a sequential machine using a network simulator, written specifically for this experiment. The network described previously was trained on a single pair of no/go utterances by a single speaker for 6000 training iterations.

The value of the squared-error term during learning was observed; it was 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 error value did reach sharply-defined minimum value after 4000 iterations; the network at that point was chosen for further study.

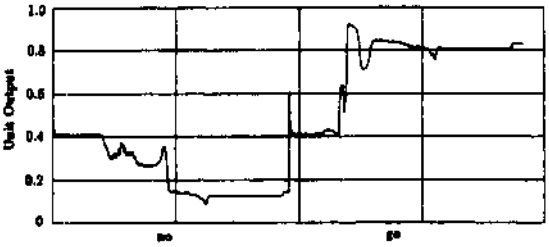

Figure 3: "Output Unit 24 Response to No/Go Pair"

3.1 Output Unit Response

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 3. 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 exhibits correct discrimination between the pair of items in the training set.

The significance of this result should not be overlooked. First, the local application of a global optimization metric provided a successful path to the desired network response pattern. Second, although no segmentation decisions were made, the network was able to form a discriminating spectral feature which was localized in time. Third, the approximations of constant weight value, and restrictions to maximum r 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.

3.2 Extension to Test Set

In order to test the generality and robustness of the internal representations obtained from the training word pair, 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 responses of the hidden units were analyzed for the 50 test utterances as well as the 2 training utterances. 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 unit failed to respond to the input data. The energy levels for this utterance were very low, especially in the mid to upper channels.

4 Discussion

Although the results of this initial experiment are unexpectedly encouraging, there are several problems which need to be addressed. The stability of the learning algorithm needs to be improved. This could be accomplished through better target functions, greater accuracy in computing the gradient, or improved learning algorithms. These ideas for improvement have been addressed in subsequent work. More powerful optimization algorithms (second-order iterative methods) have have resulted in stable learning and greatly increased learning speed.

5 Conclusions

In conclusion, several interesting results emerge from this experiment. Using a connectionist network with a temporal data flow architecture with recurrent Hnks, and using an coarse approximation of the desired output as a teaching function, a successful discriminatory mechanism was learned. This discriminatory feature was formed without segmentation and *without a direct comparison of the two items.*

The discriminatory mechanism turned out to be very robust, even though based on a single training sample. This result is very encouraging for further research with connectionist networks in deriving robust discriminatory features of phonetic classes.

Obviously, the goal of this research is to structure networks which can learn the complete set of phonetic class discriminations, so that it could support real-time, continuous speech recognition. This requires larger networks, which for efficiency, may need to be partitioned and recombined. Initial steps in this direction have been taken by training networks to discriminate the stop consonants in CV words using various vowels.

References

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

[2] 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.

[3] Jeffrey L. Elman and David Zipser. *Learning the Hidden Structure of Speech.* Technical Report ICS Report 8701, UCSD Institute for Cognitive Science, February 1987.

[4] 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.

[5] Michael I. Jordan. At tractor 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.

[6] 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.

[7] 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 Micro structure of Cognition: Volume II Psychological and Biological Models,* chapter 15, MIT Press, Cambridge, MA, 1986.

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

[9] R. W. Prager, T. D. Harrison, and F. Fallside. Boltzmann machines for speech recognition. *Computer Speech and Language,* l(l):3-27, March 1986.

[10] 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 m the Microstructure of Cognition: Volume I Foundations,* chapter 8, MIT Press, Cambridge, MA, 1986.

[11] 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.

[12] 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.

[13] 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.