LEARNED PHONETIC DISCRIMINATION USING CONNECTIONIST NETWORKS

R.L.Watrous[1] [2], L.Shastri[3], A.H.Waibel[4].

Abstract

A method for learning phonetic features from speech data using a *temporal flow model* is described, in which sampled speech data flows through a connectionist network from input to output units. The network uses hidden units with recurrent links to capture spectral/temporal characteristics of phonetic features. A simple experiment to discriminate the consonants [b,d,g] in the context of [i,a,u] using CV words is described. A supervised learning algorithm is used which performs gradient descent using a coarse approximation of the desired output as an target function. Context-dependent internal representations (features) were formed in the process of learning the discrimination task. A second experiment demonstrating learned vowel discrimination in various consonant environments is also presented. Both discrimination tasks were performed successfully without segmentation of the input, and *without a direct comparison of the training items.*

INTRODUCTION

The connectionist network approach to speech recognition is attractive because it offers a computational model 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.

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 perception of speech depends on the correct analysis of dynamic temporal/spectral relationships. The problem of designing connectionist networks which can learn these dynamic spectral/temporal characteristics has not yet been widely studied. 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.

A previous experiment showed that a simple network with recurrent links could be trained on a single instance of the word pair "no" and "go", and correctly discriminate 98% of 25 other tokens of each word for the same speaker [3]. The experiment was repeated for a second speaker and resulted in 100% discrimination performance.

An experiment is reported here which shows that connectionist networks can be optimized to discriminate the voiced stop consonants, [b,d,g], in various vowel contexts. A second experiment demonstrates the discrimination of the vowels [i,a,u] in the environment of various stop consonants. The results of these experiments show that connectionist networks can be designed and trained to successfully discriminate similar word pairs by learning context-dependent acoustic-phonetic features.

EXPERIMENT

The first experiment was designed to learn stop consonant discrimination in different vowel contexts, using CV words. The experiment used the voiced stops, [b,d,g] in three vowel contexts, [i,a,u]. A second experiment was designed to learn vowel discrimination in different consonant environments, using the same CV data.

For these experiments, a three-layer temporal flow model was implemented, as shown in Figure 1, with three output units, a variable number of hidden units, and 16 input units. The hidden and output units had self-recurrent links. The functions which define the unit behavior were chosen to approximate the computational properties of neural cells, and have convenient mathematical properties for the learning algorithm used in this experiment [2]. The unit output is a sigmoid function of the unit potential, which is the weighted sum of the outputs of the afferent units.

[1] Siemens Corp. Research, 105 College Road East, Princeton,NJ 08540
[2] Univ. of Pennsylvania, Computer and Information Sciences, Phila., PA 19104
[3] Univ. of Pennsylvania, Computer and Information Sciences
[4] ATR International Higashi-ku Osaka 540,Japan

Originally appeared in *European Conference on Speech Technology*, pp 377–380, Edinburgh (1987).

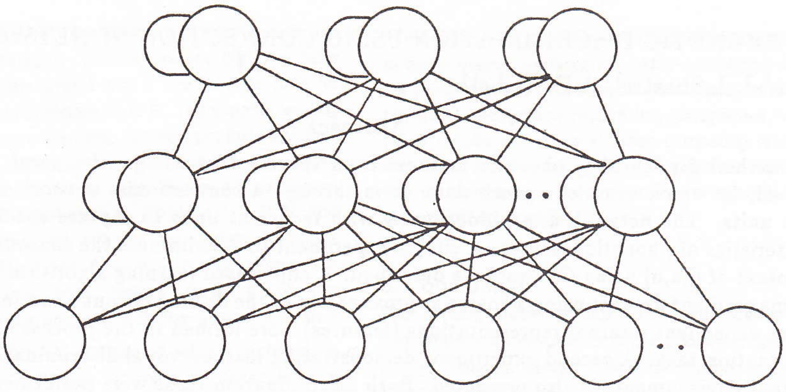

Figure 1: "Temporal Flow Model showing input, hidden and output layers"

The speech data used for these experiments consisted of isolated consonant-vowel (CV) utterances for a single speaker (RW) consisting of the stop consonants [b,d,g] in combination with the vowels [i,a,u]. Five repetitions of each CV word for a total of forty-five utterances were spoken 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 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 in the center of the vowel nucleus. This segmentation was done to decrease the computational load on the optimization algorithm and did not involve an attempt to identify the consonant-vowel boundary. It is certain that sufficient if not complete discriminatory information remained in the segmented data.

For these experiments, the Broyden-Fletcher-Goldfarb-Shanno optimization algorithm (BFGS) was used [1]. 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 function minima. The algorithm was used to modify the unit connection weights in order to minimize the mean squared error between the actual and desired output values [3].

The target function for the output units consisted of a simple Gaussian function, with a variable center point and sharpness parameter. This represented the intuition that evidence for a particular phonetic category reaches a peak near some critical point in time. For the consonant experiment, the release of the stop closure was the critical event, which occurred roughly in the center of the data buffer. For this reason the target function center value was chosen as 0.5. For the vowel experiment, the Gaussian was shifted so that the maximum was at the end of the buffer (0.9). This corresponded to the intuition that the vowel discrimination reached a maximum toward the vowel center.

The computation of the gradient vector was accomplished by an extended form of the back-propagation learning algorithm for networks with recurrent links [2,4].

A randomly initialized network with 16 hidden units was optimized for consonant discrimination. The squared error decreased from 2934 to 121 after approximately 500 iterations. The response of the output units for the optimized network can be seen in Figure 2. The output units respond in equal and opposite ways to the input stimuli; in addition, their time response roughly approximates a Gaussian. Since the learned response closely fits the training function, the network shows very good discrimination between the items of the training set. The response of the network to the other items is analogous to that shown in the figure.

The response of the hidden units to the training data was also evaluated. An example can be seen in Figure 3, where it will be noticed that the hidden unit response is decidedly context specific.

A similarly initialized network, with 10 hidden units, was optimized for vowel discrimination. The

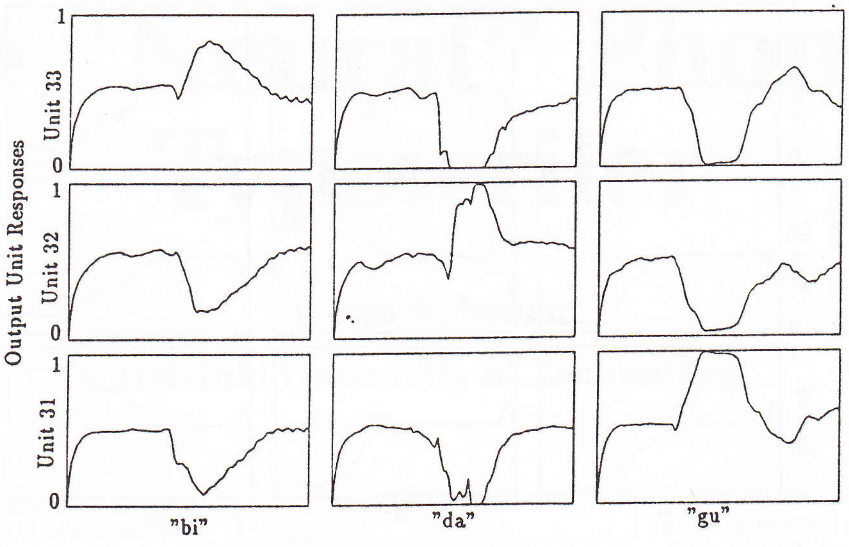

Figure 2: "Consonant Unit Responses to [bi,da,gu]"

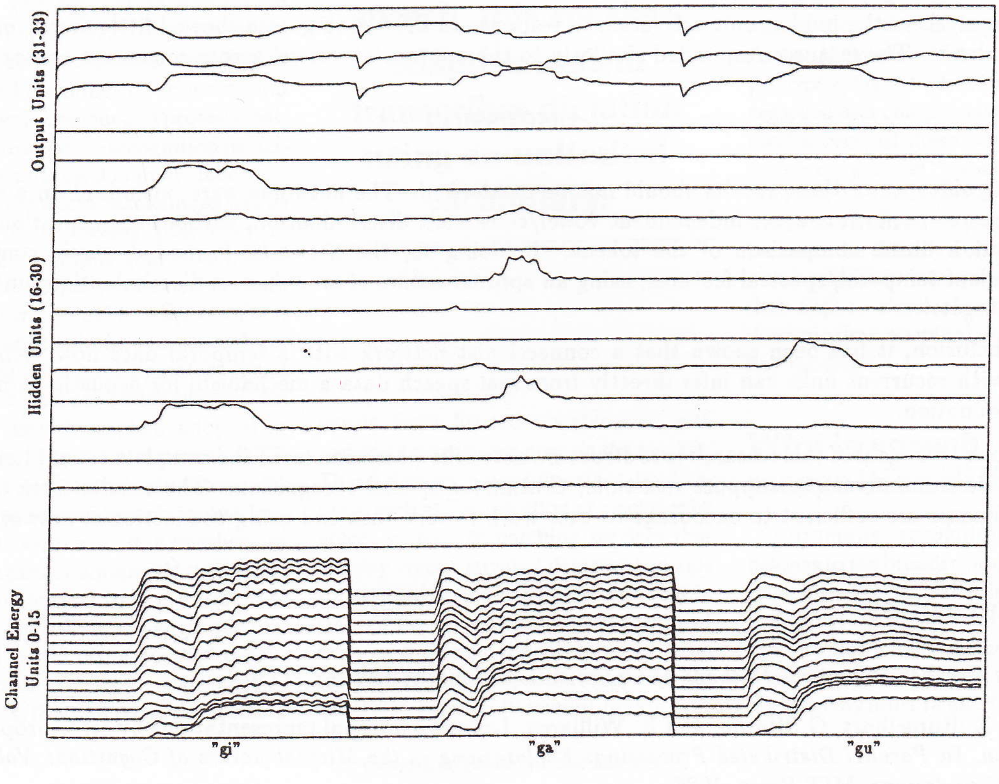

Figure 3: "Hidden Unit Responses to [gi,ga,gu]"

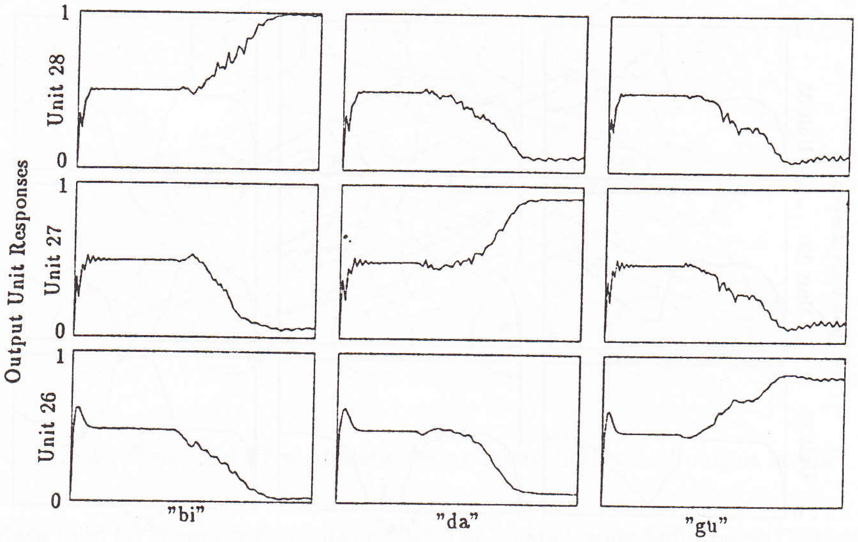

Figure 4: "Vowel Unit Responses to [bi,da,gu]"

squared error decreased from 2995 to 38.2 after approximately 140 iterations. The response of the output units for the optimized network can be seen in Figure 4. The output units show very good discrimination between the items of the training set. The response of both networks to the other items in the training set is analogous to that shown in the figures.

The analysis of the hidden unit activation in response to the training data showed little or no context dependence. The features responded similarly to the appropriate vowel across consonant contexts.

DISCUSSION

The significance of these results should not be overlooked. The networks were optimized on a small data set to perform context-independent vowel/consonant discrimination, without segmentation and without a direct comparison of the tokens. In doing so, the networks formed dynamic context-dependent temporal/spectral features, using an approximation of an unknown discrimination function as a target.

In conclusion, it has been shown that a connectionist network with a temporal data flow architecture with recurrent links can infer directly from real speech data a mechanism for acoustic-phonetic discrimination.

The long term goal of this research is to structure networks which can learn the complete set of phonetic class discriminations, to support real-time, continuous speech recognition. The results from these experiments are sufficient to encourage further work toward that end using connectionist networks.

References

[1] R. Fletcher. *Practical Methods of Optimization.* John Wiley, NY, 1980.

[2] D. E. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume I Foundations*, MIT Press, 1986.

[3] R. L. Watrous and L. Shastri. Learning acoustic features from speech data using connectionist networks. In *Proc. Cog. Sci. Conference*, July 1987.

[4] R. L. Watrous and L. Shastri. *Learning Phonetic Features Using Connectionist Networks: An Experiment in Speech Recognition.* Technical Report MS-CIS-86-78, Univ. of Penna., Oct. 1986.