Learned Classification of Sonar Targets Using a Massively Parallel Network

R. PAUL GORMAN AND TERRENCE J. SEJNOWSKI

Abstract—**We have applied massively parallel learning networks to the classification of sonar returns from two undersea targets and have studied the ability of networks to correctly classify both training and testing examples. Networks with an intermediate layer of hidden processing units achieved a classification accuracy as high as 100 percent on a training set of 104 returns. These networks correctly classified a test set of 104 returns not contained in the training set with an accuracy of up to 90.4 percent. Networks without an intermediate layer of processing units achieved only 73.1 percent correct on the same test set. Performance improved and the variability due to the initial conditions for training decreased with the number of hidden units. The effect of training set design on test set performance was also examined. The performance of a three-layered network was better than trained human listeners and the network generalized better than a nearest neighbor classifier.**

INTRODUCTION

TRADITIONAL pattern recognition techniques are often used as a first step in the interpretation of complex signals [1]–[3]. Typically, simplifying assumptions about the structure of a signal are made in order to reduce the computation required to achieve accurate classification [4]–[7]. For applications where such assumptions are valid, these techniques perform well. However, if the signals are not simply distributed or are highly correlated, these techniques may be inadequate, and other more general techniques are often impractical [8], [9].

The recent development of learning algorithms for multilayered massively parallel networks has provided potential alternatives to traditional pattern recognition which make far less restrictive assumptions about the structure of the input patterns [10], [11]. The inherent parallelism of these networks allows very rapid parallel search and best-match computations, alleviating much of the computational overhead incurred when applying traditional nonparametric techniques to signal interpretation problems. However, few studies have been conducted to determine whether such networks can learn to discriminate continuous-valued signals or to compare the performance of massively parallel networks to more traditional techniques and with human performance. (See [12]–[14] for examples of studies which begin to address these issues.)

Manuscript received February 22, 1988.

R. P. Gorman is with the Allied-Signal Aerospace Technology Center, Columbia, MD 21045.

T. J. Sejnowski is with the Department of Biophysics, The Johns Hopkins University, Baltimore, MD 21218.

IEEE Log Number 8821365.

The present study addresses the application of a massively parallel network to a signal classification problem with important practical applications, namely, the identification of undersea targets from sonar returns, and demonstrates their ability to learn to classify such complex continuous-valued signals for target discrimination. This study also addresses network performance as a function of the number of hidden units and the sensitivity of networks to initial weight values. The effect of training set design on network generalization for this particular signal classification problem is also studied.

The following two sections introduce the network architecture and the learning algorithm used in the present study. The third section discusses the classification problem and describes the network experiments. The preprocessing performed on the sonar returns for presentation to the networks is then described, followed by the experimental results. Finally, conclusions drawn from the experimental results are discussed and a comparison of the performance of the networks to the results of a previous study [15] involving trained human listeners is presented.

NETWORKS ARCHITECTURE

The networks used for the experiments discussed below were composed of three layers of processing units [Fig. 1(a)] that performed a memoryless nonlinear transformation on their summed inputs and produced continuous-valued outputs between 0.0 and 1.0. The networks were "feedforward" in the sense that each unit received input only from the units in the layer below it. Weights on connections between units were positive or negative real values. In order to determine the output of the ith unit, all of its inputs p_j were first summed as follows:

$$E_i = \sum_j w_{ij} p_j + \theta_i \tag{1}$$

where w_{ij} is the weight from the jth to the ith unit and θ_i is the bias of the ith unit. A sigmoidal transformation was then applied to the result of this summation [Fig. 1(b)].

$$p_i = P(E_i) = \frac{1}{1 + e^{-E_i}}. \tag{2}$$

The bottom (input) layer of the network was made up of 60 units, each clamped to an amplitude value of the signal to be classified. The number of output units was arbitrarily set at two. The states of output units determined the

0096-3518/88/0700-1135$01.00 © 1988 IEEE

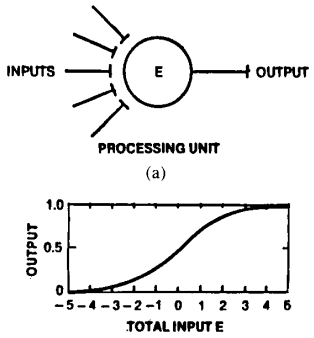


Fig. 1. (a) Schematic model of a processing unit receiving inputs from other processing units. (b) Nonlinear transformation between summed inputs and outputs of a processing unit.

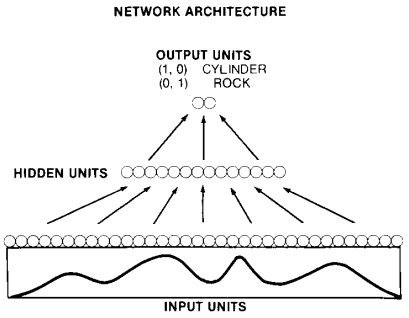


Fig. 2. Architecture of the network. The bottom layer consists of 60 processing units with their inputs "clamped" to the amplitude of the preprocessed sonar return. The hidden layer has modifiable weights on both the input and output connections, which allows the network to extract high-order features from the input waveform.

class of the signal: (1, 0) represented a return from the cylinder, and (0, 1) represented a return from the rock. An intermediate or "hidden" layer, which allows the network to extract high-order correlations in the signal, transformed the input pattern to the appropriate output pattern. A schematic of the basic architecture is shown in Fig. 2.

Network Learning Algorithm

The backpropagation learning algorithm [11] was used to train the network. For each learning cycle, the input layer was initially "clamped" to a sample waveform from the training set. This simply means that the output of the ith unit in the input layer was assigned the value of the ith waveform value. The waveform was normalized to the range 0.0–1.0. The activity of each unit was propagated forward through each layer of the network using (1) and (2). The activity at the output layer was compared to the desired activity, and an error $\delta_i^{(N)}$ for each output unit was calculated as follows:

$$\delta_i^{(N)} = (p_i^* - p_i)P'(E_i^{(N)}) \tag{3}$$

where N is the number of layers in the network, p_i is the activity of the output unit, and p_i^* is the desired activity. $P'(\ )$ is the first derivative of $P(\ )$. The error at the output was then backpropagated recursively to each lower layer (n) as follows:

$$\delta_i^{(n)} = \sum_j \delta_j^{(n+1)} w_{ij}^{(n)} P'(E_i^{(n)}) \tag{4}$$

where $w_{ij}^{(n)}$ is the weight from the jth unit in layer n to the ith unit in layer $n + 1$. This error was backpropagated only when the difference between the measured and desired activities at the output unit was greater than a margin of 0.2. In order for the network to learn, the value of each weight had to be incrementally adjusted in proportion to the contribution of each unit to the total error. The change in each weight was calculated as follows:

$$\Delta w_{ij}^{(n)} = \epsilon \delta_i^{(n+1)} p_j^{(n)} \tag{5}$$

where ϵ controls the rate of learning (a value of 2.0 was used for these experiments). The weights of the network were initialized to small random values uniformly distributed between -0.3 and 0.3. This was done to prevent the hidden units from acquiring identical weights during training. The networks were simulated on a Ridge 32 computer (comparable to a VAX 780 in computational power) using a simulator written in the C programming language and developed at The Johns Hopkins University.

Sonar Data

The classification problem addressed in this study was undersea target identification. Target echos from an active sonar system had to be classified as a return from the appropriate target. The data used for the network experiments were sonar returns collected from a metal cylinder and a cylindrically shaped rock positioned lengthwise on a sandy ocean floor. The impinging pulse was a wideband linear FM chirp. Returns were obtained from each target at various aspect angles.

A set of 208 returns (111 cylinder returns and 97 rock returns) were selected on the basis of the strength of the specular return (4.0–15.0 dB signal-to-noise ratio), making certain that a variety aspect angles were represented. The processed representation used as input to the network was chosen as the result of other experiments with human listeners [15]. First, a short-term Fourier transform $F(t, \nu)$ of the sonar return $f(t)$ was obtained:

$$F(t, \nu) = \int_{t-T/2}^{t+T/2} f(\tau) e^{-i\nu\tau}\, d\tau \tag{6}$$

where T is the width of each temporal segment. From $F(t, \nu)$, the spectral envelope $P_{t_0,\nu_0}(\eta)$ was computed:

$$P_{t_0,\nu_0}(\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |F(t, \nu)|^2 \omega\left(t - \left(t_0 + \frac{\Delta t}{2} + \eta\sigma_t\right), \nu - \left(\nu_0 + \frac{\Delta\nu}{2} + \eta\sigma_f\right)\right) dt\, d\nu$$
$$\eta = 0, 1, 2 \cdots \eta_{\max} \tag{7}$$

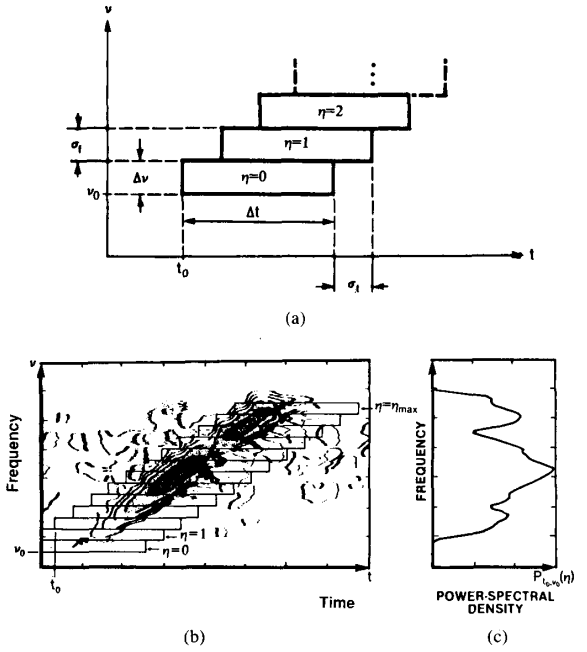


Fig. 3. The preprocessing of the sonar return produces a sampled spectral envelope normalized to vary from 0.0 to 1.0 for input to the network. (a) The set of sampling apertures offset temporally to correspond to the slope of the FM chirp. (b) Sampling apertures superimposed over the 2-D display of the short-term Fourier transform. (c) The spectral envelope obtained by integrating over each sampling aperture.

where ν_0 and t_0 are the starting frequency and temporal position of the FM chirp, respectively. The sampling aperture ω, with temporal and spectral dimensions given by Δt and $\Delta\nu$, respectively, was defined as

$$\omega(t, \nu) = 1 \begin{cases} -\frac{\Delta t}{2} < t < \frac{\Delta t}{2} \\ -\frac{\Delta \nu}{2} < \nu < \frac{\Delta \nu}{2} \end{cases}$$
$$= 0 \quad \text{otherwise.} \tag{8}$$

The $\eta_{\max}$ locations of the aperture were determined by the temporal and spectral sampling intervals $\sigma_t = 0.1\Delta t$ and $\sigma_f = \Delta\nu$. The values of σ_t and σ_f are related by

$$\frac{\sigma_f}{\sigma_t} = \frac{d\nu}{dt} \tag{9}$$

where $d\nu/dt$ is the slope of the FM chirp. This process is indicated schematically in Fig. 3 where a set of sampling apertures are superimposed over the 2-D display of the short-term Fourier transform spectrogram of the sonar return. As shown in Fig. 3(b) and (c), the spectral envelope $P_{t_0,\nu_0}(\eta)$ was obtained by integrating over each aperture. Sixty sample points were obtained for each envelope. These samples were normalized to take on values between 0.0 and 1.0 and were used as input to the network.

Network Experiments

Experiments were designed to address several issues. The central issue was to determine whether a network could be trained to classify the targets by presenting the network with examples of sonar returns from each target. In addition, two factors affecting network performance were examined, namely, the number of hidden units and the initial weight values. Finally, the performance of the network on the test set is compared for networks trained on randomly selected training sets (an aspect-angle independent experiment), and a training set selected to contain examples from each aspect angle represented in the total set of sonar returns (an aspect-angle dependent experiment). This comparison allowed us to evaluate the impact of aspect-angle dependent information on classification accuracy.

For both the aspect-angle independent and aspect-angle dependent experiments, a training set was selected from the total set of returns, and performance, in terms of percent correct classification, was determined after 300 presentations of the training set to the learning network. Generalization was tested by measuring the performance of trained networks on the set of returns excluded from the training set.

To determine the sensitivity to initial conditions, each experiment was repeated ten times with different, randomly selected, initial weight values. This procedure was used to train networks with varying numbers of hidden units in order to evaluate the role of the hidden layer in network performance. Networks with 0, 2, 3, 6, 12, and 24 hidden units were trained and tested using identical training and testing sets and their performance was compared.

For the aspect-angle independent experiment, the above paradigm was repeated 13 times using different training sets. For each iteration, a set of 16 returns were randomly selected from the total set of 208 returns to serve as the test set, and the remaining 192 returns were used to train the networks. Testing sets were selected so that each sonar return served only once as a test signal. This allowed 13 experiments to be conducted with disjoint testing sets.

This procedure is routinely used in pattern recognition to obtain a more accurate estimate of the probability of misclassification given a finite set of samples [16]. The variation in performance across test sets provided an indication that features required to accurately classify a signal varied from one return to the next. This suggested that some important signal patterns were aspect-angle dependent.

For the aspect-angle dependent experiment, the training and testing set were designed to ensure that returns from each target aspect angle represented in the total data set were included with representative frequency in both the training and the testing set. Both the training and the testing set consisted of 104 returns. The networks' performance was again taken as the average performance over ten trials.

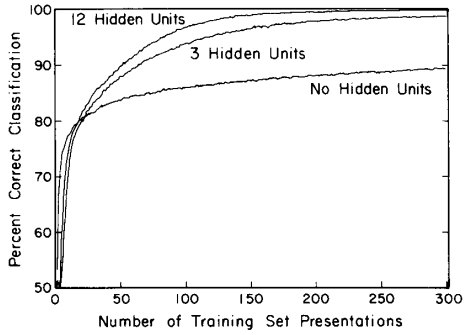


Fig. 4. Network learning curves for experiments with randomly chosen training sets. Each curve represents an average of 130 learning trials for a network with the specified number of hidden units.

TABLE I
SUMMARY OF THE RESULTS OF THE EXPERIMENT WITH RANDOMLY SELECTED TRAINING SETS. THE STANDARD DEVIATION SHOWN IS ACROSS TRAINING AND TESTING SETS, AND WAS OBTAINED BY MEASURING THE VARIATION OF PERFORMANCE VALUES AVERAGED OVER TEN TRIALS DIFFERING IN INITIAL CONDITIONS

Number of Hidden Units	Average Performance on Training Sets (percent)	Standard Deviation on Training Sets (percent)	Average Performance on Testing Sets (percent)	Standard Deviation on Testing Sets (percent)
0	89.4	2.1	77.1	8.3
2	96.5	0.7	81.9	6.2
3	98.8	0.4	82.0	7.3
6	99.7	0.2	83.5	5.6
12	99.8	0.1	84.7	5.7
24	99.8	0.1	84.5	5.7

EXPERIMENTAL RESULTS

The aspect-angle independent experiment conducted using randomly selected training sets consisted of 130 trials for each network with a given number of hidden units. The overall performance of each network was taken to be the average over a set of 13 values obtained from experiments with different training sets. These 13 values were in turn averaged over ten trials differing in initial conditions. Fig. 4 shows the overall average learning curves for three of the networks trained on randomly selected returns. The best average performance was achieved by a network with 24 hidden units (99.8 percent correct classification accuracy). The network with no hidden units, essentially an Adaline [17], could classify the training set with an average accuracy of 89.4 percent, indicating that, on average, the hidden layer contributed at least 10 percent to the networks' performance. However, because of the ceiling effect on the performance of networks with large hidden layers, it may be a low estimate of the importance of the hidden layer in general.

The results of this set of experiments are summarized in Table I. The standard deviation reported is the variation over 13 average performance values. Each average performance value was obtained over the ten trials differing in initial conditions. Thus, this variation is primarily due to training set selection. Average performance on the training and testing sets improved with the number of hidden units up to 12 hidden units. Increasing the number of hidden units from 12 to 24 produced no further improvement.

For the aspect-angle dependent experiment, the average learning curves for the networks were similar to those for the aspect-angle independent experiment shown in Fig. 4, with the best performance of 100 percent being attained by the network with 24 units. The two-layered network achieved an accuracy of only 79.3 percent on this training set, 10 percent lower than the first experiment, whereas the performance of the networks with hidden units was slightly higher. The performance of the two-layered network on the test set was also lower in the second set of experiments (73.1 percent compared to 77.1 percent), while the performance of the networks with hidden units was markedly better.

The results of the second experiment are summarized in Table II. The variation reported for this experiment was only over ten trials differing in the initial conditions. Again, the performance increased with the number of hidden units up to 12 units. In addition, the variation in performance of networks with hidden units decreased as the number of hidden units increased.

TABLE II
SUMMARY OF THE RESULTS OF THE EXPERIMENT WITH TRAINING AND TESTING SETS SELECTED TO INCLUDE ALL TARGET ASPECT ANGLES. THE STANDARD DEVIATION SHOWN IS ACROSS NETWORKS WITH DIFFERENT INITIAL CONDITIONS

Number of Hidden Units	Average Performance on Training Sets (percent)	Standard Deviation on Training Sets (percent)	Average Performance on Testing Sets (percent)	Standard Deviation on Testing Sets (percent)
0	79.3	3.4	73.1	4.8
2	96.2	2.2	85.7	6.3
3	98.1	1.5	87.6	3.0
6	99.4	0.9	89.3	2.4
12	99.8	0.6	90.4	1.8
24	100.0	0.0	89.2	1.4

DISCUSSION

Massively parallel networks have been trained to identify two undersea targets on the basis of single sonar returns. Two experiments were conducted. In an aspect-angle independent experiment, training and testing sets were selected at random and, in an aspect-angle dependent experiment, these returns were selected to ensure that all target angles in the total set of sonar returns were represented in both the training and testing sets. In both experiments, the networks with hidden units could be trained to achieve a high degree of classification accuracy.

The performance of the two-layered network (without hidden units) was 10–20 percent lower than the three-layered networks on the training sets and 8–17 percent lower on the testing sets, which supports previous findings on the importance of the hidden layer for difficult classification problems [18], [19]. The performance of the networks also tended to improve as the number hidden units was increased from 0 to 12, but little or no improvement was realized with the increase from 12 to 24 units.

When the number of training samples is small compared to the number of adjustable weights, the information capacity of the network may exceed the total amount of information contained in the set of samples. In such cases, the network would not be sufficiently constrained to force generalization, and therefore might simply memorize each individual pattern in the training set. Such a solution would prove of little value for the general target recognition problem.

In the present study, the number of adjustable weights generally exceeded the number of training samples. However, the improvement in performance with the number of hidden units suggests that the capacity of the networks did not exceed the information contained in the training set. Furthermore, the performance of trained networks on test waveforms not contained in the training set demonstrates that these networks utilized general signal features to achieve accurate classification. It is interesting to note that the performance on the testing sets did not deteriorate when the number of hidden units was increased from 12 to 24, which effectively doubled the number of weights. Apparently, the extra degrees of freedom did not reduce the ability of the network to generalize from a limited number of training examples as occurs in other problems [20].

The variation in performance due to initial conditions was moderate for networks with few or no hidden units, and decreased with increasing numbers of hidden units. This suggests that networks with larger hidden layers tend to be less sensitive to initial conditions. The variance on the test set shown in Table I is higher than the variation shown in Table II because the variation in performance in the first experiment included an additional factor due to the choice of training and testing examples.

The performance of the three-layered networks on the test sets was higher in the aspect-angle dependent experiment than in the aspect-angle independent experiment. This supports the interpretation of the variation in performance on test sets in the aspect-angle independent experiment as being due to the exclusion of important aspect-angle dependent patterns from the randomly selected training sets. This implies that certain features extracted by the networks were related to target geometries. However, the consistent achievement of between 80 and 84 percent performance by three-layered networks, independent of the set of training examples, suggests that aspect-angle independent information, such as target material, may have also been important.

The lower performance of the two-layered network on the training and testing set in the second experiment as compared to the first may have been as the result of a number of experimental factors. The total number of input pattern presentations was higher in the aspect-angle independent experiment due to the larger number of examples in the training set. The number of distinct patterns in the aspect-angle dependent experiment may have also been greater due to the inclusion of all aspect angles in the training set. Finally, the fact that performance was very near perfect for three-layered networks may have distorted the comparative performance between two- and three-layered networks.

The performance of the network classifier can be compared, in some respects, to human performance and to the performance of a linear classifier based on human perceptual features. In a previous experiment [15], three human

subjects were trained to classify the same two targets by listening to single returns taken from the set of returns used to train the networks. The human listeners were presented with the raw signal data shifted down to auditory frequencies. They continued to train until their performance ceased to improve. The average performance of the three trained human listeners was 91 percent. The performance of the networks, trained on preprocessed versions of the signals, was close to 100 percent.

The comparison to trained listeners is incomplete in several respects. First, the humans were not tested for generalization, although these experiments are planned. Second, different signal representations were used for the human listeners and the networks, and therefore, different information may have been utilized for signal classification. An analysis of trained networks to determine the signal features used would perhaps address this issue. Nonetheless, sonar target recognition is one area where human performance has been consistently better than automatic systems, and the superior performance of massively parallel networks on a similar task suggests that such networks may provide a viable alternative to current techniques.

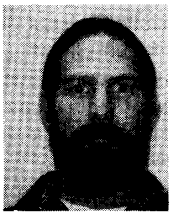
Finally, a nearest neighbor classifier ($k = 1$) was developed for comparative purposes. The classification of each return was determined by the class of its nearest neighbor according to a Euclidean metric. The performance of this classifier was 82 percent correct classification on the total set of 208 returns. The performance of network classifiers with hidden units on test sets was consistently better than the nearest neighbor classifier.

Acknowledgment

The authors wish to thank R. Burne, D. Goblirsch, G. Hinton, D. Mitchell, T. Sawatari, and R. Simpson for many insightful discussions during the course of this work. The network simulator used in the present study was based on programs written by P. Kienker and C. Rosenberg.

References

[1] S. E. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE*, vol. 73, pp. 1625-1650, Nov. 1985.
[2] C. H. Chen, "On statistical and structural feature selection," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed. New York: Academic, 1976, pp. 135-144.
[3] K. S. Fu, "Hybrid approaches to pattern recognition," in *Pattern Recognition Theory and Applications*, J. Kittler, K. S. Fu, and L. F. Pau, Eds. Boston, MA: Reidel, 1982, pp. 139-156.
[4] ——, "Statistical pattern recognition," in *Adaptive, Learning, and Pattern Recognition: Theory and Applications*, J. M. Mendel and K. S. Fu, Eds. New York: Academic, 1970, pp. 35-79.
[5] L. Kanal, "Patterns in pattern recognition: 1968-1974," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 697-719, Nov. 1974.
[6] T. M. Cover, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21-27, Jan. 1967.
[7] K. S. Fu, P. J. Min, and T. J. Li, "Feature selection in pattern recognition," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-6, pp. 33-27, Jan. 1970.
[8] K. S. Fu, "Recent developments in pattern recognition," *IEEE Trans. Comput.*, vol. C-29, pp. 845-854, Oct. 1980.
[9] R. M. Haralick, "Decision making in context," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 417-428, July 1983.
[10] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Sci.*, vol. 9, pp. 147-169, 1985.
[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: M.I.T. Press, 1986, pp. 318-362.
[12] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," ATR Tech. Rep. TR-1-0006, ATR Interpreting Telephony Res. Lab., 1987.
[13] R. L. Watrous and L. Shastri, "Learning acoustic features from speech data using connectionist networks," in *Proc. 9th Annu. Conf. Cognitive Sci. Soc.*, July 1987, pp. 518-530.
[14] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, Apr. 1987.
[15] R. P. Gorman and T. Sawatari, "Automatic sonar target recognition based on human perceptual features," submitted to *J. Acoust. Soc. Amer.*, July 1987.
[16] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 472-479, July 1974.
[17] G. Widrow and M. E. Hoff, "Adaptive switching circuits," in *Proc. IRE Western Electron. Show Conv.*, part 4, 1960, pp. 96-104.
[18] T. J. Sejnowski, P. K. Kienker, and G. H. Hinton, "Learning symmetry groups with hidden units: Beyond the perceptron," *Phys. D*, pp. 260-275, 1986.
[19] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex Syst.*, vol. 1, pp. 145-168, 1987.
[20] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, "Large automatic learning, rule extraction, and generalization," *Complex Syst.*, vol. 1, pp. 877-922, Oct. 1987.



R. Paul Gorman received the B.S. degree in physics from Wayne State University, Detroit, MI, in 1981 and is currently a graduate student in electrical engineering at The Johns Hopkins University, Baltimore, MD, specializing in massively parallel computation.

He joined the Bendix Research Laboratories in 1978 as a Research Technician. In 1981 he joined the Bendix Advanced Technology Center as an Associate Member of the Technical Staff. Currently, as a member of the Technical Staff at the Allied-Signal Aerospace Technology Center, he is involved in the development of massively parallel adaptive architectures for signal processing and sensor-based target recognition.

Terrence J. Sejnowski received the B.S. degree in physics from Case Western Reserve University, Cleveland, OH, and the Ph.D. degree in physics from Princeton University, Princeton, NJ, in 1978.

He was a Postdoctoral Fellow in the Department of Neurobiology at Harvard School until 1982, when he joined the Faculty in the Department of Biophysics at Johns Hopkins University. He is currently a Professor of Biophysics, Biology, Computer Science, and Electrical and Computer Engineering. His main research interest is in the representation, transformation, and storage of information in the nervous system.

Dr. Sejnowski received a Presidential Young Investigator Award in 1984 and was a Wiersma Visiting Professor of Neurobiology at the California Institute of Technology in 1987.