Neural Networks, Vol. 1, pp. 75–89, 1988
Printed in the USA.

0893-6080/88 $3.00 + .00

ORIGINAL CONTRIBUTION

Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets

R. PAUL GORMAN

Allied-Signal Aerospace Technology Center

TERRENCE J. SEJNOWSKI

Johns Hopkins University

(*Received and accepted* 30 *October* 1987)

Abstract—*A neural network learning procedure has been applied to the classification of sonar returns from two undersea targets, a metal cylinder and a similarly shaped rock. Networks with an intermediate layer of hidden processing units achieved a classification accuracy as high as 100% on a training set of 104 returns. These networks correctly classified up to 90.4% of 104 test returns not contained in the training set. This performance was better than that of a nearest neighbor classifier, which was 82.7%, and was close to that of an optimal Bayes classifier. Specific signal features extracted by hidden units in a trained network were identified and related to coding schemes in the pattern of connection strengths between the input and the hidden units. Network performance and classification strategy was comparable to that of trained human listeners.*

Keywords—Learning algorithms, Hidden units, Multilayered neural network, Sonar, Signal processing.

INTRODUCTION

New learning algorithms for multilayered neural networks are potential alternatives to existing pattern recognition and signal processing techniques (Lapedes & Farber, 1987; Lippman, 1987; Watrous & Shastri, 1987). The application of neural networks to signal classification problems requires far less restrictive assumptions about the structure of the input signal. In addition, 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 non-parametric techniques to signal interpretation problems. However, these new techniques are not yet fully characterized and it would be valuable to compare their performance with that of existing techniques and human performance on a well-defined problem.

In the present study, neural networks were applied to a sonar target classification problem. Networks were trained to classify sonar returns from an undersea metal cylinder and a cylindrically shaped rock of comparable size. Some of the performance data were reported elsewhere (Gorman & Sejnowski, 1987). In this paper, we present additional experimental data, a comparison of network classification performance to that of traditional pattern recognition techniques, and an analysis of the networks' classification strategy. We also compare the features extracted by the network's hidden units to perceptual cues used by trained human listeners.

The following section discusses the network architecture and the learning algorithm used in the present study. The preprocessing performed on the sonar returns for presentation to the networks is described in the second section. The third and fourth sections describe the classification experiments and present experimental results. The fifth section discusses a technique for analyzing network weight patterns and its application to a network trained to classify the two sonar targets. Finally, a comparison is drawn between

The authors wish to thank David Goblirsch, Donald Mitchell, and Takeo Sawatari for their contribution to the present study. We also thank Richard Burne, Moise Goldstein, Geoffrey Hinton, Ning Qian, and Robert Simpson for many insightful discussions during the course of this work. Finally, we thank Patrick Coleman and Paul Payer for their assistance as trained human listeners. The network simulator used in the present study was based on programs written by Paul Kienker and Charles Rosenberg.

Requests for reprints should be sent to Dr. Terrence J. Sejnowski, Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218.

the network classifier and human listeners trained to discriminate the same two classes of sonar returns. The paper concludes with a brief discussion of the results.

NETWORK ARCHITECTURE AND LEARNING ALGORITHM

The networks used in the present study were feed-forward and possessed two or three layers of processing units with continuous-valued outputs. The output of the ith unit was obtained by calculating the activation level, E_i,

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

where w_{ij} is the weight from the jth to the ith unit, p_j is the output of unit j, and b_i is the bias of the ith unit. A sigmoidal transformation was then applied to the activation level to obtain the ith unit's state or output p_i,

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

where β was a constant that determined the slope of the sigmoid at $E_i = 0$ ($\beta = 1.0$ for these experiments). The input layer 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 the output units determined the class of the signal: (1,0) represented a return from the metal cylinder, and (0,1) represented a return from the rock. Experiments were conducted using networks with two layers and networks with a hidden layer. A schematic of the three-layered architecture is shown in Figure 1.

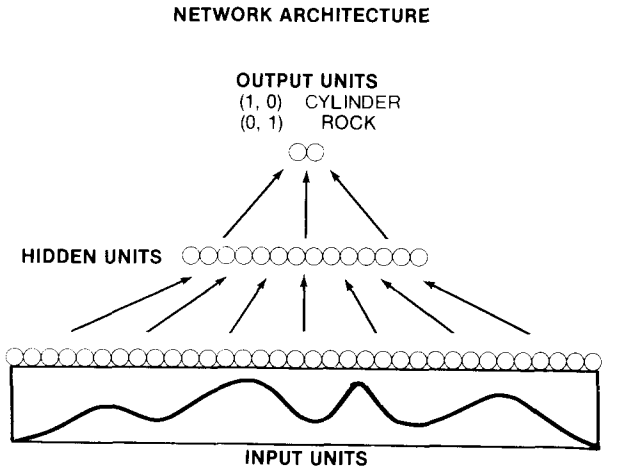

FIGURE 1. Schematic diagram of the network. The bottom layer has 60 processing units with their states "clamped" to the amplitude of the pre-processed sonar signal, shown in analog form below the units. The two output units at the top represent the two sonar targets to be identified. The layer of hidden units between the input and output layers allows the network to extract high-order signal features. The connections between the layers of units are represented by arrows.

The back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986) was used to train the network. The algorithm calculated the gradient of the error with respect to each weight in the network and incrementally adjusted the weights to minimize the global error (see Gorman & Sejnowski, 1987 and Rosenberg & Sejnowski, 1987 for more details). The error measured at each output unit was back-propagated only when the difference between the measured and desired states of the output unit was greater than a margin of 0.2. 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. In all experiments, the learning rate parameter, ϵ, was set to 2.0 and momentum, α, was 0.0, as defined in Rosenberg and Sejnowski (1987). The networks were simulated on a Ridge 32 computer (comparable to a VAX 780 FPA in computational power) using a simulator written in the C programming language and developed at The Johns Hopkins University.

SONAR DATA AND SIGNAL REPRESENTATION

The data used for the network experiments were sonar returns collected from a metal cylinder and a cylindrically shaped rock positioned on a sandy ocean floor. Both targets were approximately 5 ft in length and the impinging pulse was a wide-band linear FM chirp ($ka = 55.6$). Returns were collected at a range of 10 meters and obtained from the cylinder at aspect angles spanning 90° and from the rock at aspect angles spanning 180°.

A set of 208 returns (111 cylinder returns and 97 rock returns) were selected from a total set of 1200 returns on the basis of the strength of the specular return (4.0 to 15.0 dB signal-to-noise ratio). An average of 5 returns were selected from each aspect angle. Figure 2 shows a sample return from the rock and the cylinder. The preprocessing of the raw signal was based on experiments with human listeners (Gorman & Sawatari, 1987). The temporal signal was first filtered and spectral information was extracted and used to represent the signal on the input layer.

The preprocessing used to obtain the spectral envelope is indicated schematically in Figure 3 where a set of sampling apertures (Figure 3a) are superimposed over the 2D display of a short-term Fourier Transform spectrogram of the sonar return. As shown in Figure 3b and c, the spectral envelope, $P_{t_0,v_0}(\eta)$, was obtained by integrating over each aperture. The spectral envelope was composed of 60 spectral samples, normalized to take on values between 0.0 and 1.0. (See Gorman & Sejnowski, 1987 for a detailed treatment of the preprocessing).

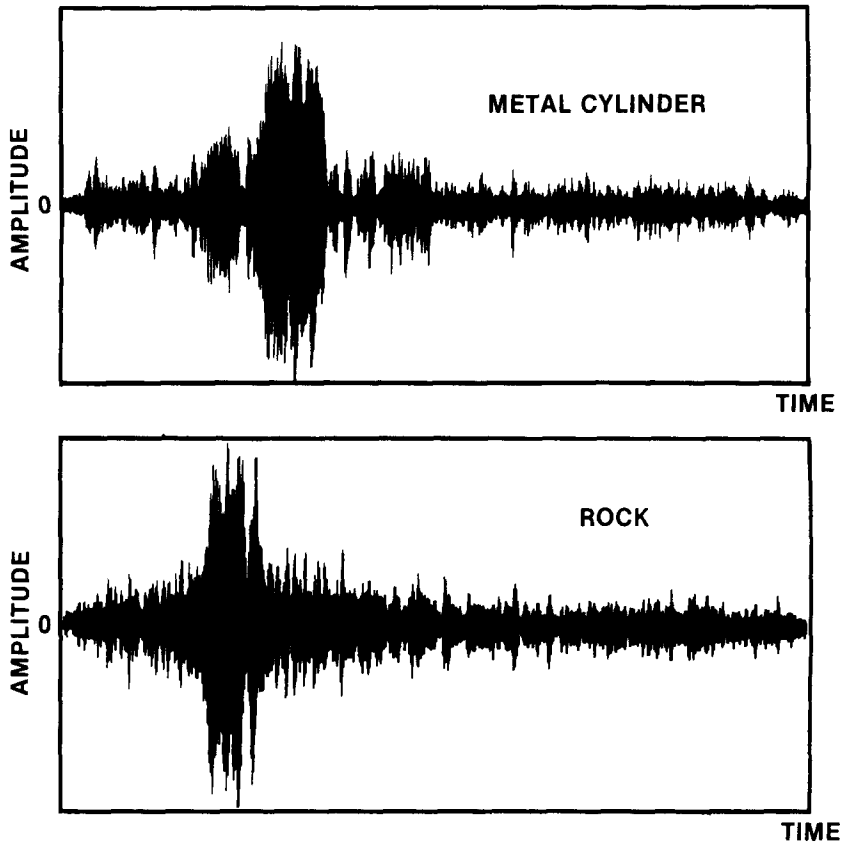

FIGURE 2. Amplitude displays of a typical return from the cylinder and the rock as a function of time.

CLASSIFICATION EXPERIMENTS

Parallel networks were trained to classify returns from the rock and the cylinder according to the procedure outlined above. For each experiment, a given network was presented with a sequence of training signals during which weight values were changed to improve performance. The training set was presented a total of 300 times. The trained network was then presented with a set of test returns excluded from the training set to determine its ability to generalize. The network's performance on both the training and test set was specified as the percent correct classification. Each experiment with a given network was repeated 10 times with different initial weight values to average over variations in performance due to initial conditions.

Two series of experiments were conducted: an aspect-angle independent series for which training sets were selected at random from the total set of 208 returns, and an aspect-angle dependent series using training and testing sets designed to contain examples from all available aspect-angles. This permitted us to compare the test-set performance of trained networks in these two series to determine whether aspect-angle dependent signal features were important for accurate classfication.

Aspect-Angle Independent Series

For the aspect-angle independent series, 13 disjoint test sets comprised of 16 returns were randomly selected. The 192 returns remaining after each test set selection served as the corresponding training set. In this way, each signal in the total set of 208 served as a testing signal for one of the experiments. A given network was trained and tested on each training/testing set pair and the average performance over the 13 test sets provided a measure of the probability of correct classification (Toussaint, 1974). This measure would be accurate as long as no important classification features were excluded from any of the training sets; otherwise, the measure would be lower than the expected performance. Network performance on each of the 13 training/testing set pairs was computed as the average performance of 10 separately trained networks.

Aspect-Angle Dependent Series

For the aspect-angle dependent series, the training and testing sets were designed to ensure that both sets contained returns from each target aspect angle with representative frequency. Each set consisted of 104 returns. The networks' performance was again taken as the average performance over ten separately trained networks.

Number of Hidden Units

The number of hidden units required to accurately classify the returns was determined empirically. Both

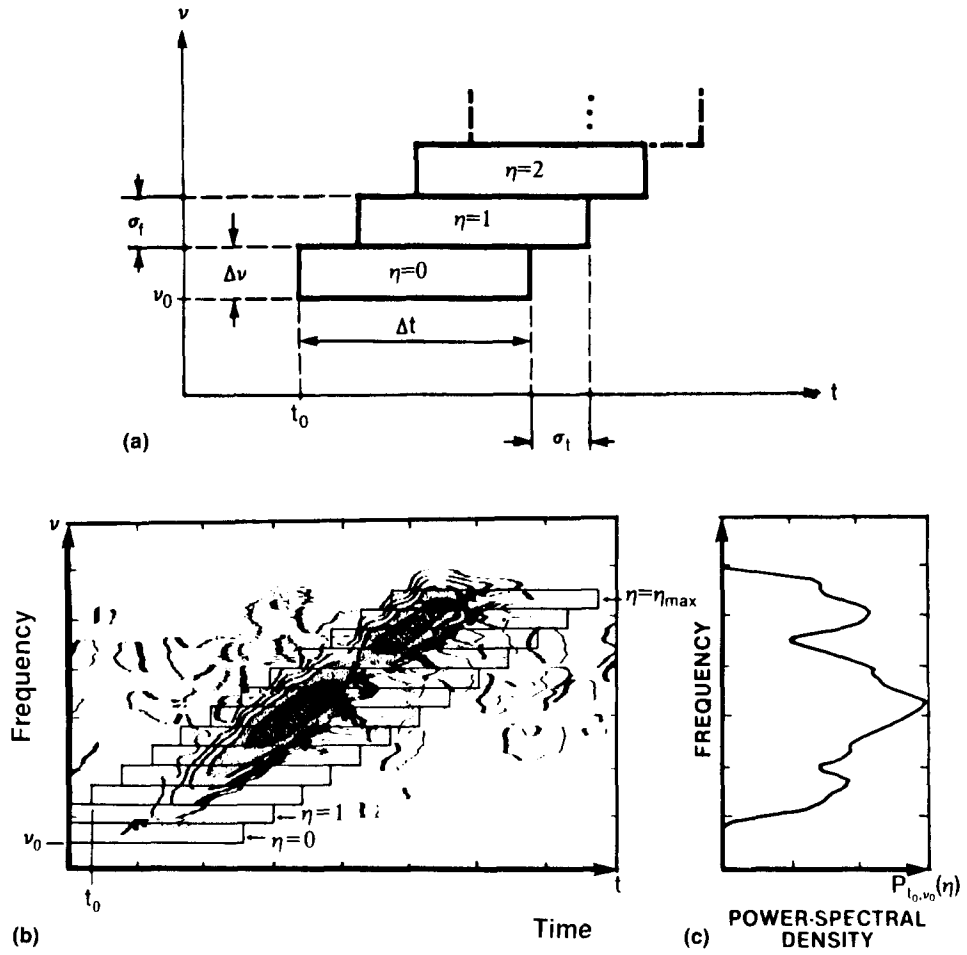

FIGURE 3. The preprocessing of the sonar signal produces a sampled spectral envelope normalized to vary from 0.0 to 1.0 for input to the network. (a) The set of sampling apertures offset temporally to correspond to the slope of the FM chirp, (b) sampling apertures superimposed over the 2D display of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

of the above series of experiments were repeated using networks with 0, 2, 3, 6, 12, and 24 hidden units.

Nearest Neighbor Classification

The performance of our network classifier was compared with a nearest neighbor classifier which served as a performance benchmark. A nearest neighbor rule classifies an unknown signal in the category of its nearest neighbor according to a pre-specified measure of similarity or distance. We used a Euclidean metric so that the distance, d_{kl}, between the kth and lth signal was defined as

$$d_{kl} = \left[\sum_{i=1}^{N} (x_i^{(k)} - x_i^{(l)})^2\right]^{1/2} \tag{3}$$

where N is the number of spectral envelope sample points, and $x_i^{(k)}$ is the value of the ith sample point for the kth signal.

The probability of correct classification was based on computing the nearest neighbor to each signal in the database. If the two nearest neighbor signals were returns from the same target the classifier was given a score of 1.0. If the two signals were returns from different targets the classifier was given a score of 0.0. The probability of correct classification is given by

$$P(c) = \frac{1}{M}\sum_{k=1}^{M} s_k \tag{4}$$

where M is the number of returns and s_k is the classification score of the kth signal. The probability of correct classification of the nearest neighbor classifier can be used to obtain upper and lower bounds on the Bayes probability of correct classification as the number of sample signals increases (Cover & Hart, 1967). Hence, the performance of the nearest neighbor classifier provides a reasonably good benchmark for rating the performance of a network classifier.

EXPERIMENTAL RESULTS

Aspect-Angle Independent Series

The aspect-angle independent series conducted using randomly selected training sets consisted of 130 trials

TABLE 1
Aspect-Angle Independent Series

Number of Hidden Units	Average Performance on Training Sets (%)	Standard Deviation on Training Sets (%)	Average Performance on Testing Sets (%)	Standard Deviation on Testing Sets (%)
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

Summary of the results of the aspect-angle independent series of experiments 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.

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 averages over 10 trials differing in initial conditions. The results of this series of experiments are summarized in Table 1. Figure 4 shows the overall average learning curves for three of the networks trained on randomly selected returns.

The best average performance on the training set was achieved by a network with 24 hidden units (99.8% correct classification accuracy). The network with no hidden units, essentially an Adaline (Widrow & Hoff, 1960), could classify the training set with an average accuracy of 89.4%, indicating that, on average, the hidden layer improved the networks' performance by at least 10%.

The average performance on the test set differed by as much as 7.6% between two- and three-layered networks. The 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.

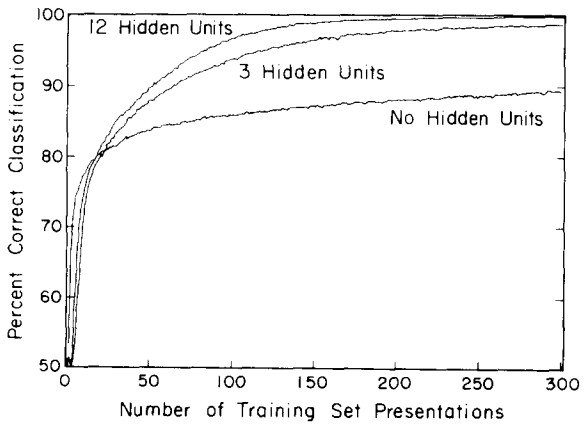

FIGURE 4. Network learning curves for the aspect-angle independent series of experiments using randomly chosen training sets. Each curve represents an average of 130 learning trials for a network with the specified number of hidden units.

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.

The amount of variation in performance, as reported in Table 1, suggests that the increase in performance with the size of the hidden layer is not very significant. However, the way that the performance of each network varied as function of test set was correlated. This can be seen in Figure 5 which shows a plot of the performance of four of the networks tested as a function of the test set. The performance of the network generally increased as a function of the size of the hidden layer, particularly on test sets for which performance was low.

Aspect-Angle Dependent Series

The results of the aspect-angle dependent experiments are summarized in Table 2. The average learning curves for the this series is shown in Figure 6. The best performance of 100% was attained by the network with 24 hidden units. The two-layered network achieved an accuracy of only 79.3% on this training set, 10% lower than in the first series of experiments, whereas the performance of the networks with hidden units remained the same. The performance of the two-layered network on the test set was also lower in the second set of experiments (73.1% compared to 77.1%), while the performance of the networks with hidden units was as much as 5.7% better.

The variation reported for this experiment was only over ten trials differing in the initial conditions. Again, the performance on the test set increased with the number of hidden units up to 12 units and the variation in performance of networks with hidden units decreased as the number of hidden units increased.

The performance of networks with hidden units on the test set in the aspect-angle dependent series was consistently better then the test set performance of the same networks in the aspect-angle independent series.

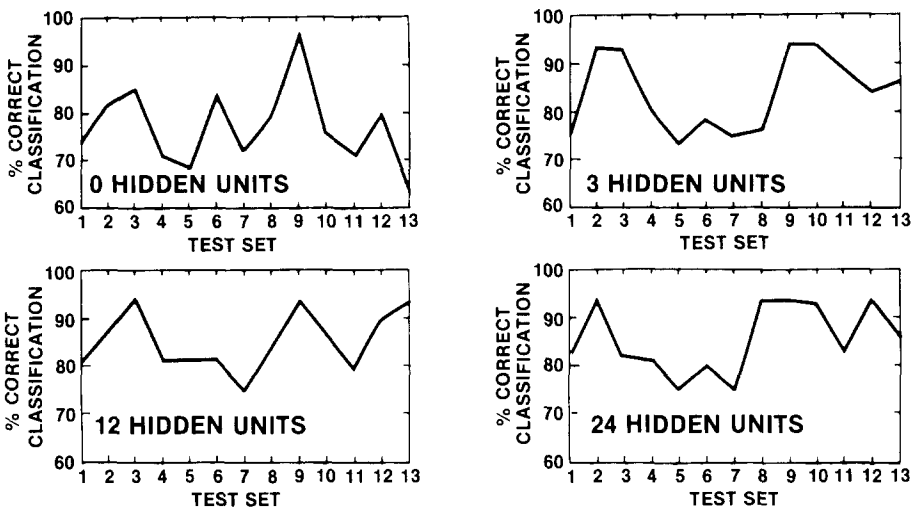

FIGURE 5. Network performance as a function of test set selected randomly for the aspect-angle independent series of experiments. Results are shown for networks with 0, 3, 12, and 24 hidden units. The test-set sequence was the same for each hidden unit series.

The low average performance in the independent series can be attributed to particular training/testing set pairs for which the performance dropped below 80% (see Figure 5). In these instances, important signal patterns appeared in the testing sets but were not represented in the training set and were misclassified. If correct classification were dependent solely upon general patterns associated with each class and represented in each sample signal, then the variation in performance across training/testing set pairs should not be observed. This variation indicates that patterns associated with specific aspect-angles are also important for accurate classification.

Effect of Hidden Units on Network Performance

The significance of the observed increase in network performance with the number of hidden units was tested by an analysis of variance on the results of the aspect-angle dependent test experiments. Using an F distribution, we verified that varying the number of hidden units did create a difference in test performance, $F(5, 54) = 32.3$, $p < .001$. Networks with hidden units performed better than those without hidden units, $F(1, 54) = 147.29$, $p < .001$, and it was advantageous to use six or more hidden units, $F(1, 54) = 12.02$, $p < .01$. In addition, there was a linear component in the relation between performance and the number of hidden units meaning that performance did tend to increase as the number of hidden units increased, $F(1, 54) = 52.7$, $p < .001$. Finally, there was also a quadratic component indicating that there may be a peak performance as a function of the number of hidden units, $F(1, 54) = 65.94$, $p < .001$ although this was not observed up to an asymptotic limit of 24 hidden units.

Nearest Neighbor Classifier

The probability of correct classification using a nearest neighbor rule computed from Equation 4 was 82.7%. This was as much as 1.7% lower than the performance of the network classifier with hidden units in

TABLE 2
Aspect-Angle Dependent Series

Number of Hidden Units	Average Performance on Training Sets (%)	Standard Deviation on Training Sets (%)	Average Performance on Testing Sets (%)	Standard Deviation on Testing Sets (%)
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

Summary of the results of the aspect-angle dependent series of experiments with training and testing sets selected to include all target aspect angles. The standard deviation shown is across networks with different initial conditions.

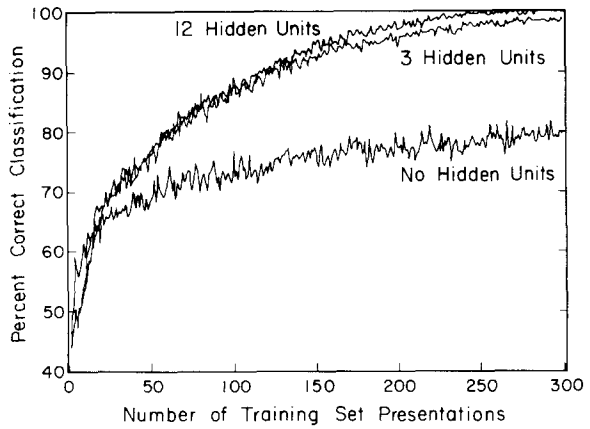

FIGURE 6. Network learning curves for the aspect-angle dependent series of experiments. Each curve represents an average of 10 learning trials for a network with the specified number of hidden units.

the aspect-angle independent series. If we assume that this result is an accurate estimate of the performance of the nearest neighbor classifier in the limit as the number of samples increases, the performance of a Bayes classifier, given the underlying probability structure of the signals, would lie between 82.7% and 91.4% (Cover & Hart, 1967). The performance of the network classifier with 12 hidden units in the aspect-angle dependent series was near the top of this range. This suggests that the performance of the network classifier was near optimal.

WEIGHT PATTERN ANALYSIS

Method of Analysis

In addition to demonstrating the ability of neural networks to classify complex signals, we were interested in understanding the classification strategy discovered by the networks. One way to accomplish this is to interpret the patterns of weights on connections between processing units in trained networks that are matched to structure in the signals. We chose a trained network with only three hidden units, but with good performance, to simplify the analysis.

Figure 7 shows the pattern of weights of a network with three hidden units trained to classify returns from the rock and cylinder. Each hidden unit is represented in the figure by a labeled grey region. The area of the rectangles within these regions is proportional to the absolute value of the weights on connections to other units in the network. White and black rectangles represent positive and negative weights, respectively. The lower set of 60 rectangles represent weights on connections from input units. The two upper rectangles in the center represent weights to the output units and the single rectangle in the upper left represents the weight from the true unit or bias to the hidden unit.

The classification strategy of the network cannot be readily understood by visually inspecting weight displays. Different signal patterns will interact with the weight patterns in different ways, so the signals themselves must also be included in the analysis. In particular, it is important to characterize the signals that produce the highest activation for each hidden unit. This is analogous to the concept of best feature for sensory neurons in the nervous system. The first step was to analyze each hidden unit independently and then to determine how the hidden units interact to achieve accurate classification. A set of signal patterns was obtained by clustering the set of sample signals using a weighted metric that depended on the weights on connections from the input units to the hidden unit being analyzed. To analyze the ith hidden unit, a weight-state vector $Q_j^{(k)}[i]$ for each sample signal, k was first computed

$$Q_j^{(k)}[i] = [w_{ij} p_j^{(k)}] \tag{5}$$

where $p_j^{(k)}$ is the output of the jth input unit when the kth signal is clamped to the input. Figure 8 shows a graphic representation of the weight-state vector.

The Euclidean distance between each pair of weight-state vectors was computed using Equation 3. A hierarchical clustering technique (Johnson, 1967) was then

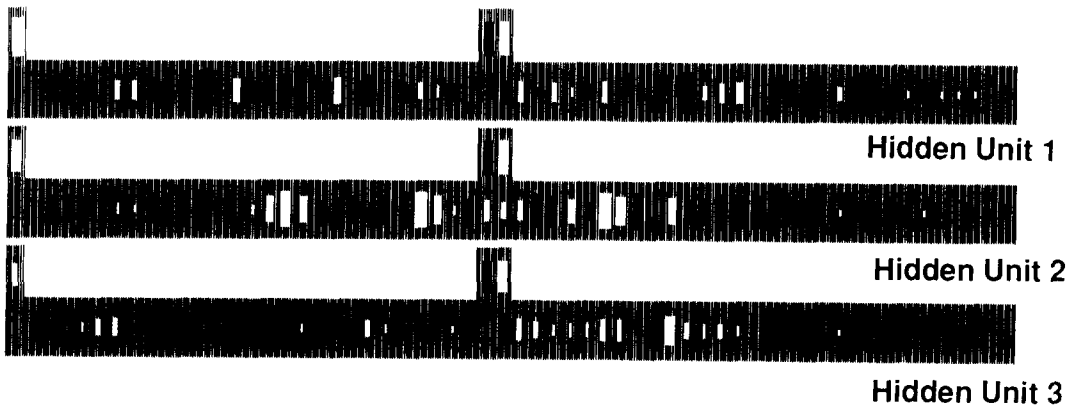

FIGURE 7. The weight pattern for a trained network with three hidden units. See text for explanation.

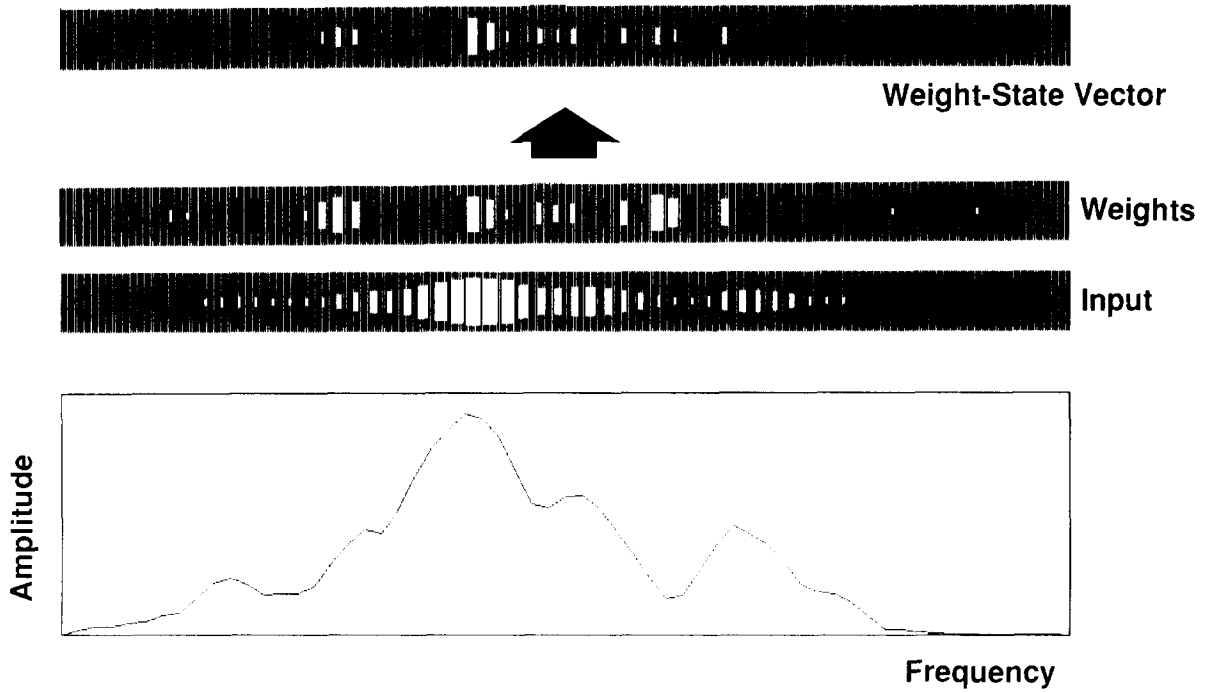

FIGURE 8. The derivation of the weight-state vector from the product of weight values on the input to each hidden unit and output values of input units when clamped to a given sonar return. A plot of the spectral envelope is shown at the bottom and the output of input units is represented in the lower gray region. The middle gray region represents the weights on connections from the input units to a given hidden unit. The product of the output values and the weights produces the weight-state vector represented by the top gray region.

applied to this distance matrix. Each cluster was a set of sonar returns whose weight-state vectors were similar to each other. The cluster centroids, computed by averaging the signal vectors over all members of each cluster, comprised a set of distinct patterns which could be ordered along the dimension defined by the hidden unit response. This dimension was interpreted as a signal feature that the hidden unit learned to extract.

Analysis of a Trained Network

A network with three hidden units, trained in the aspect-angle independent series on 192 sonar returns was chosen for analysis. The network's performance on the training set and testing set was 100% and 93.7%, respectively. Thus, this network was able to correctly classify all but one of the total set of 208 returns. Twenty-one signal clusters were obtained by analyzing the center hidden unit of the network. The weight pattern of this network is shown in Figure 7. These clusters are graphically represented in Figure 9.

The cluster boundaries were defined such that returns within each cluster produced a similar hidden unit response. The number of signals per cluster varied from 4 to 22. The centroid of each cluster was computed and the activation level response of the hidden unit to each centroid was obtained by applying Equation 1. The centroids were rank ordered according to the hidden unit activation level from most inhibitory to most excitatory, which ranged from −23.0 to 5.2. Cylinder patterns inhibited the unit, while rock patterns excited the unit. The response to 9 of the centroids, representing 50% of the returns, ranged between −3.0 and −6.5. This would indicate that the hidden unit responded strongly to about half of the input signals.

Figure 10 shows a sequence of 9 cluster centroids, representing 50% of the input signals, which span the range of hidden unit activation levels. The variation of these patterns as a function of hidden unit activation level can be characterized in terms of three signal features. As the overall bandwidth of the pattern decreases, the unit's activation level increases. Also, the onset and decay characteristics of the signals change as a function of activation level. Wide-band signals tend to have a gradual overall onset and decay while narrow-band signals do not.

The pattern of weights, shown in Figure 7, are appropriate for encoding these signal features. The weights at the extremes of the input array are sufficiently inhibitory to turn off the hidden unit when wide-band signals are clamped to the input. In addition, the alternating bands of positive and negative weights code for rate of onset and decay. If the onset or decay of the pattern is sufficiently gradual, the signal energy spans the positive and negative bands, shutting down the units response. If the rate is rapid and the band is positioned appropriately, the activation will not be balanced and will result in a net positive activation.

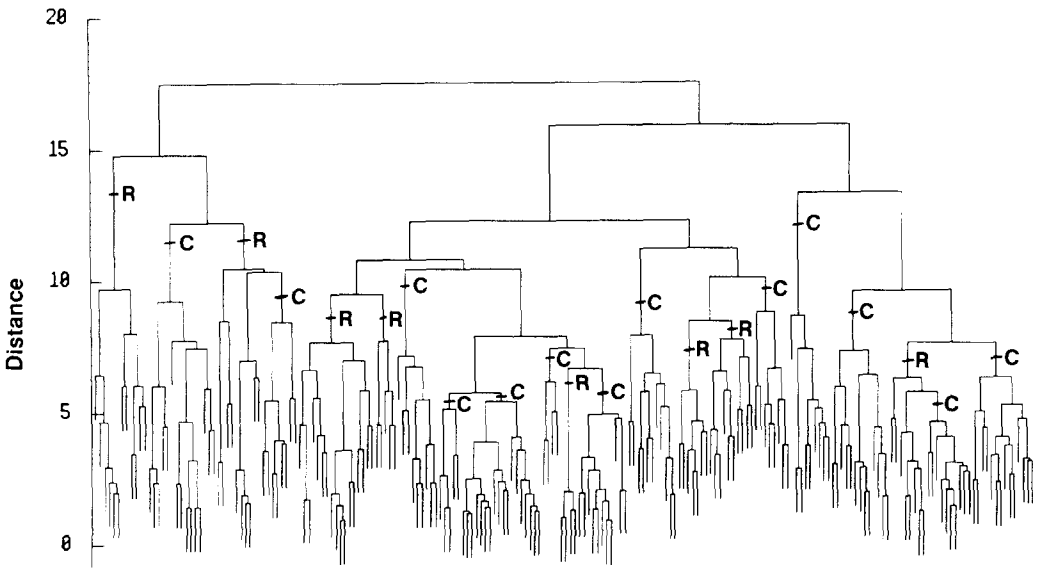

FIGURE 9. A graphical representation of the clustered sonar returns obtained by applying the weight-state clustering technique to the center hidden unit of the network shown in Figure 7. Twenty-one clusters are indicated by "C" (cylinder) or "R" (rock) located at the lowest node superior to all returns contained in the cluster. Thus, each return belongs to one and only one cluster. The cluster boundaries were determined by discontinuities in hidden unit activation levels.

These signal features represent signal dimensions that characterize variations in the input patterns to which the hidden unit has been adaptively tuned, allowing it to successfully classify about 50% of the signals presented to it. These features are general in the sense that they are aspect-angle independent. This hidden unit was also able to classify an additional 20% of the input patterns that did not fit this trend. These exceptions were represented by five centroids one of which is shown in Figure 11. This display is similar to Figure 7 except that the area of the inner rectangles are now proportional to the weight-state product. The input pattern is plotted below the network. This signal appears to be a wide-band rock which runs counter to the general description provided above. Yet this signal is correctly classified by the center hidden unit as indicated by the plot of activation level at the upper-left of the unit and the weight-state product values at the ouput.

The correct classification of such exceptional signals is achieved by coding for the precise locations of spectral peaks and nulls in the signal by positive and negative weights, respectively, between the input units and the hidden units. The opposite strategy would apply to peaks and nulls of exceptional cylinder returns. The location of these peaks could be precisely encoded by a few weights suggesting that these features were aspect-angle dependent. This is consistent with the fact that these clusters represented only 4 to 8 returns.

The use of such specific pattern information allows the hidden units to "memorize" less frequent patterns. By comparison, the general strategy outline above was encoded through the use of many weights, characterizing the essential features of a wide variety of input patterns. This suggests that the network's generalization strategy is dependent upon the relative frequency and the stability of features in the input pattern. This also explains the difference in test-set performance between the aspect-angle dependent and aspect-angle independent series of experiments, since 20% of the signals were classified correctly on the basis of aspect-angle specific information carried by only a few signals.

The remaining 30% of the signals for which the above hidden unit was poorly tuned were handled by the cooperative coding of the remaining hidden units. By repeating the above analysis for each hidden unit, it was found that the differential response to narrow-band and wide-band signals as well as onset and decay characteristics was a general strategy of all three hidden units. The response to narrow-band signals differed among the hidden units in that each was tuned to detect a narrow-band signal with a different central frequency. This strategy made it possible for our registration-sensitive network to encode a shift-invariant signal feature.

This behavior can be seen in Figure 12. The three displays depict the response of the network to three narrow-band signals with different central frequencies. In the top display only the center unit responds to the input signal. The other two units are inhibited. In the center display a signal's band is more central and the top unit responds while the other two are inhibited. Finally, the bottom display shows a narrow-band signal with a high central frequency and only the lower unit responds. These hidden units could also capture exceptions to this rule using coding schemes outlined above.

The general behavior of each hidden unit could then be used to determine the strategy of the network as a whole, as illustrated in Figure 13. A cylinder return

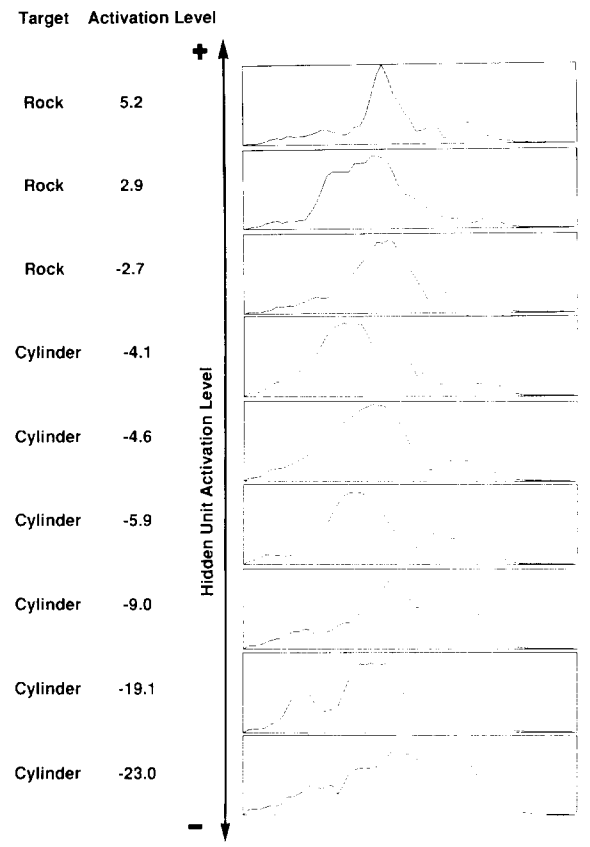

FIGURE 10. Plots of cluster centroids rank ordered according to the hidden unit activation level generated by applying the centroid pattern to the input of the trained network. The variation in spectral pattern with hidden unit response can be characterized by bandwidth, onset, and decay features.

will typically turn off all of the hidden units (Figure 13a). The default response of the network is determined by the biases to the output units. With no activation from the hidden layer, the bias drives the output layer to code for a cylinder. But a weak response from any one of the hidden units is sufficient to flip the states of the output layer to code for a rock (Figure 13b).

COMPARISON WITH HUMAN PERFORMANCE

In a previous study (Gorman & Sawatari, 1987), human subjects were trained to discriminate between the same two targets by listening to the same set of sonar returns as used in the present study. Each training exercise required the subject to listen and respond to a set of 100 returns. Single returns were presented in random order so that information about the way the returns changed with aspect angle could not be used to classify the returns. Subjects were immediately told whether they had classified a return correctly or incorrectly. The training regimen began with returns that were easy to distinguish and gradually included more difficult returns as training proceeded.

The subjects were trained on returns that were heterodyned versions of the original temporal signal. The processed FM chirp was 250 ms in duration and swept from 500 Hz to 1100 Hz. The spectral envelope used as input to the networks was extracted from this processed signal. The total duration of each stimulus was 400 ms which included the shadow region of the return (see Figure 14). Important classification cues used by human listeners were contained in the shadow region of some returns. This portion of the return was not included in the signal representation used as input to the networks.

The performance of three trained human subjects on a given set of 100 returns selected randomly from the total set of 208 training samples ranged from 88% to 97% correct. We recently trained one new subject on the same training set used to train networks in the aspect-angle dependent series. Figure 15 shows the training curve for this subject. The best performance achieved by the subject was 88%. The subject's performance on the test set used in the aspect-angle dependent series was 82%. The performance of networks with 12 hidden units was better than the performance of human listeners trained on the same set of data.

In the same previous study, the trained human listeners were tested to determine the perceptual cues used to discriminate between the two classes of returns. Verbal labels for these perceptual features were established and described in qualitative terms by the subjects. The subjects were asked to rate each return in terms of the prominence of each feature which provided a quantitative measure of the relative strength of features in each return. Two of the features identified by subjects were labeled "attack" and "decay." The attack feature was associated with the onset of cylinder returns, and decay was associated with the end of rock returns. The ratings of sonar returns along the "attack" perceptual dimension provided by subjects correlated well with a measure of the area under the low-frequency portion of the spectral envelope. The ratings along the "decay" dimension correlated well with a linear combination of the half-power spectral bandwidth and the area under the high-frequency portion of the spectral envelope (see Figure 16).

As discussed in the section on weight pattern analysis, the network classifier also extracted features related to the bandwidth and the onset and decay characteristics of the spectral envelope. If the features used by human listeners and the features extracted by the learning networks had a common signal correlate, we would expect returns rated by human subjects as having a prominent cylinder attack to inhibit hidden units in trained networks. Conversely, we would expect returns rated as possessing a strong rock decay to excite or turn on at least one hidden unit in trained networks.

We tested this hypothesis by comparing the human perceptual ratings with the responses of hidden units

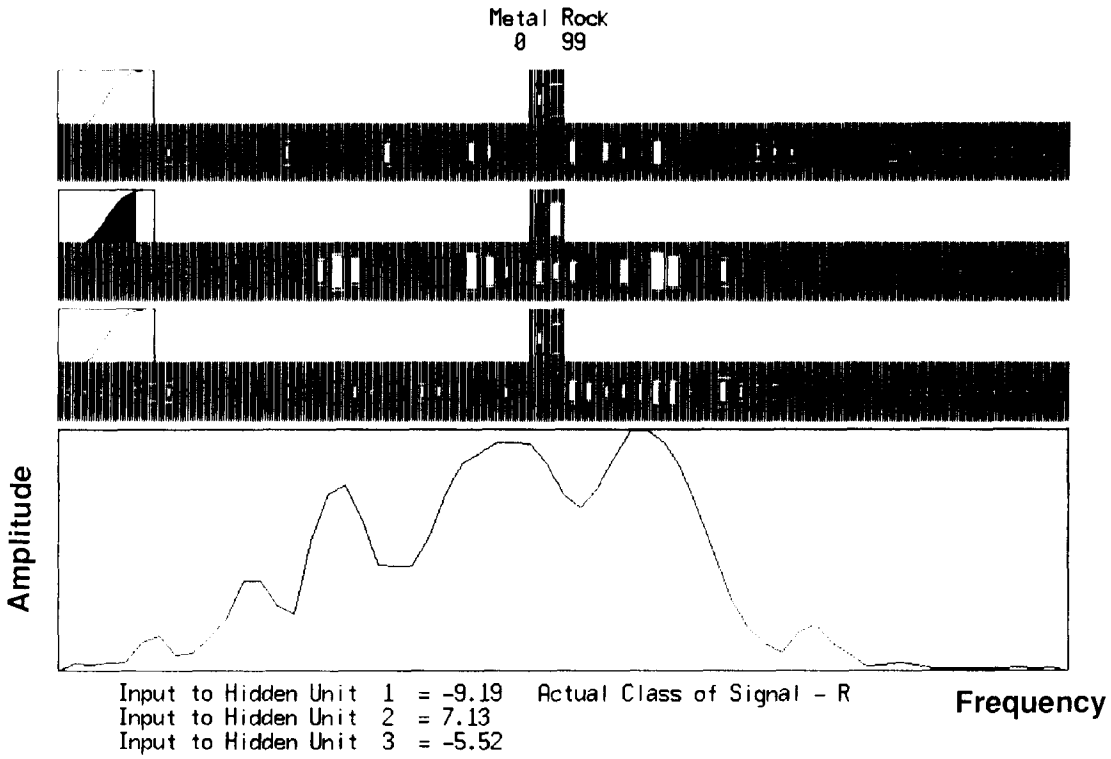

FIGURE 11. The classification of exceptional patterns requires the precise coding of specific spectral peak and null locations in the input signal. This wide-band rock return is correctly classified because of the excitatory response to the two peaks flanking the central peak.

inside a trained network. We generated a single scale from the average attack and decay ratings of humans by normalizing attack ratings to −1.0 for the most prominent cylinder attack and 0.0 for the least prominent and normalizing the decay ratings to 1.0 for the most prominent rock decay and 0.0 for the least prominent. In addition, we used the trained network analyzed in the previous section to obtain comparable network ratings for the same set of returns. The measure chosen to serve as a network rating was the output of the hidden units. For a given return, the hidden unit with the largest absolute activation level provided the rating. This allowed the hidden unit best tuned to the input to rate the return.

A product-moment correlation coefficient was computed between these normalized subject ratings and the activation levels of the appropriate hidden unit. The coefficient obtained across 100 returns was 0.84. The high correlation between these measures suggests that the network discovered features in the sonar returns that were similar to features used by trained human listeners.

DISCUSSION

Neural networks have been trained to identify two undersea targets, a metal cylinder and a similarly shaped rock, on the basis of single sonar returns. Two series of experiments were conducted. For the aspect-angle independent series, training and testing sets were selected at random, and for the aspect-angle dependent experiment, these returns were selected to ensure that all target aspect 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.

Classification Experiments

The results of the network classification experiments, as well as the analysis of variance, demonstrate that the hidden layer contributed signficantly to the performance of the network classifier. This supports previous findings on the importance of the hidden layer for difficult signal classification and signal processing problems (Lapedes & Farber, 1987; Lehky & Sejnowski, 1987; Rosenberg & Sejnowski, 1987; Sejnowski, Kienker, & Hinton, 1986). The analysis of variance indicated that a peak performance might be attained as the number of hidden units increased.

The best performance of the network classifiers with hidden units was better than a nearest neighbor classifier and performance in the aspect-angle dependent series was close to the Bayes classifier, which is an optimal decision rule for maximizing the probability of correct classification. The performance of the networks with hidden units in the aspect-angle dependent series

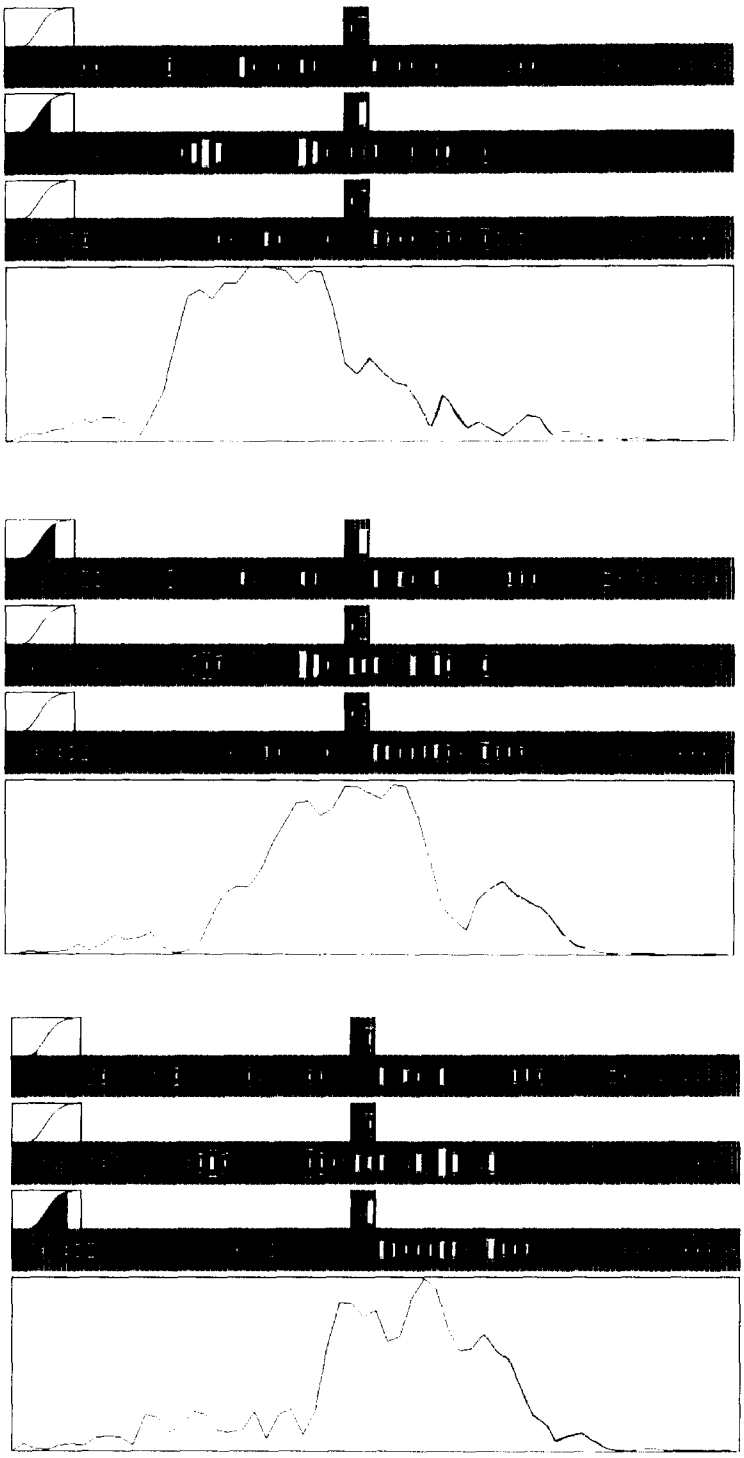

FIGURE 12. The response of hidden units to narrow-band input patterns. Each grey bar represents a hidden unit and the area of the rectangles within are proportional to the weight-state product of each input unit. White rectangles are excitatory and black rectangles are inhibitory. The sigmoids at the left of each hidden unit show the output state of the hidden unit as a function of the activation level. As indicated, hidden units respond preferentially to narrow-band patterns with different central frequencies.

was as much as 5.7% better than in the aspect-angle independent series. This indicates that aspect-angle dependent signal features contributed to the test performance of the network classifiers.

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 1 is higher then

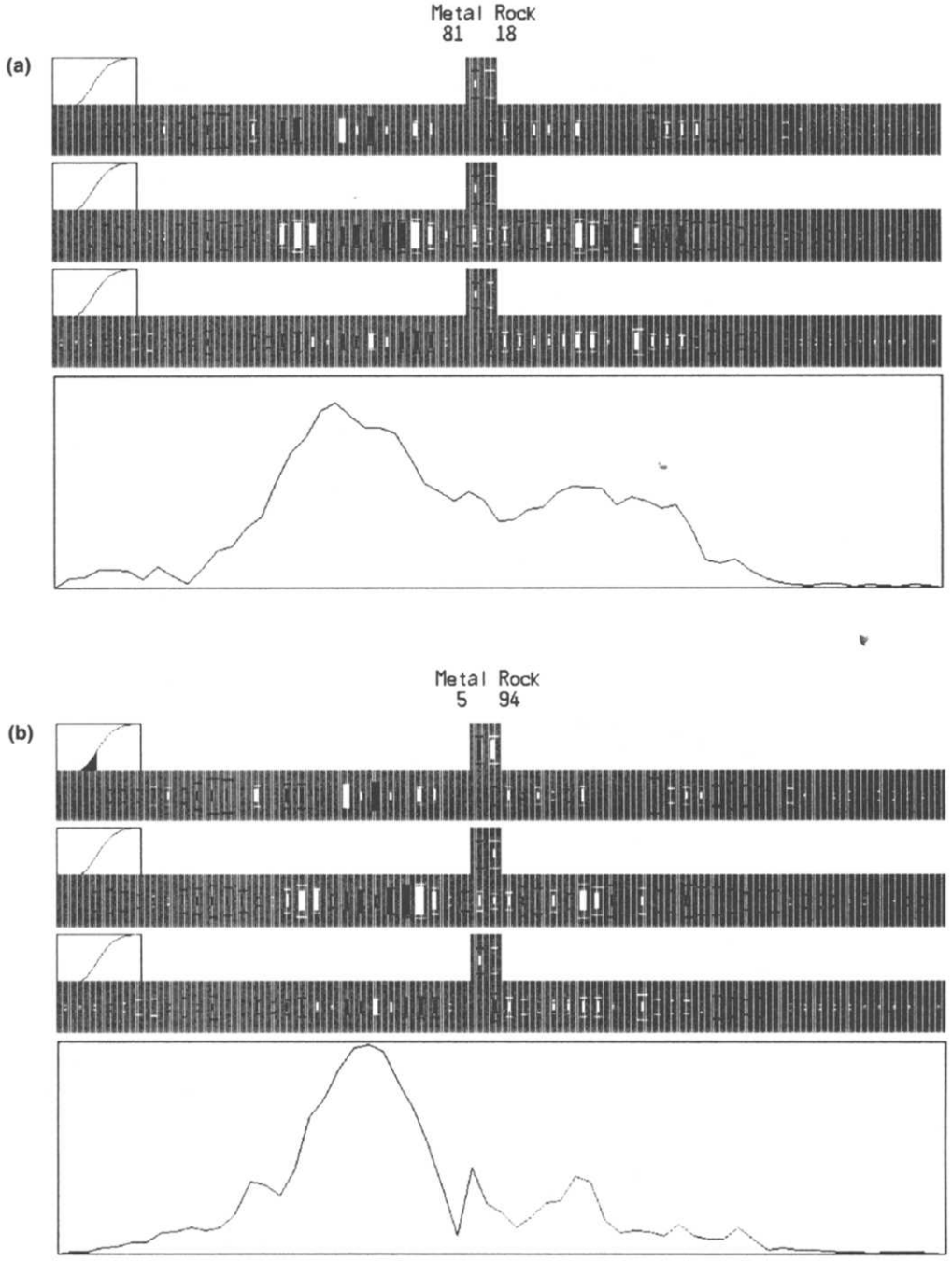

FIGURE 13. (a) The general strategy of the trained network is to default to a cylinder using the input from the bias unit, (b) only weak responses of hidden units are required to code for rock returns.

the variation shown in Table 2 because the variation in performance in the first experiment included an additional factor due to the choice of training and testing examples.

Weight Pattern Analysis

The analysis of the weight patterns contributed significantly to our understanding of the network classifier. Specific signal features extracted by hidden units in a trained network were identified and related to coding schemes in the pattern of input weights. Previous attempts at analyzing weight patterns were aided either by input patterns whose structure could be easily interpreted visually (Lehky & Sejnowski, 1987; Sejnowski *et al.*, 1986), symbolic input patterns whose primitives were given a priori (Rosenberg & Sejnowski, 1987), or a complete mathematical description of the input data (Lapedes & Farber, 1987). For signal-based recognition applications, such information about the underlying

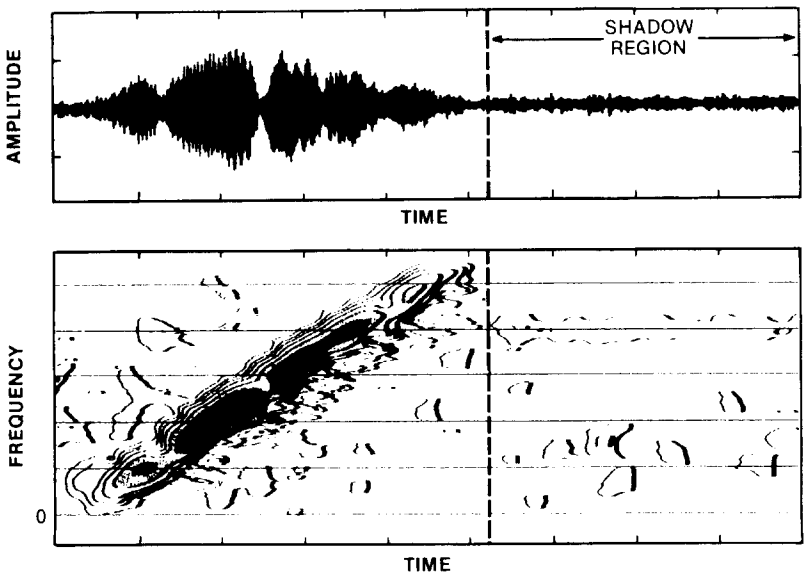

FIGURE 14. The temporal and spectrographic displays of a typical cylinder return. The "shadow region" marked in the display contained important classification cues used by human listeners. This portion of the return was not presented to the networks.

structure of the input data is not generally available. For these applications the weight-state clustering technique could aid in the interpretation of the network weight patterns.

The overall network strategy was to default to a cylinder response, and to detect the presence of a rock return at the input. Once the response characteristics of each hidden unit were understood, the conditions underlying this general strategy became apparent. One of the aspect-angle independent features of rock returns was a narrow-band input pattern. This was a shift-invariant feature since the central frequency of the band could vary. However, a single hidden unit could detect this feature only for a limited range of central frequencies. Thus, multiple hidden units responding to different ranges of central frequencies were required to encode this feature reliably. The broad-band pattern of the cylinder return, on the other hand, could be rejected with the same coding scheme by each hidden unit. The cylinder could thus serve as a default even though there was a wide range of variability among returns from the cylinder.

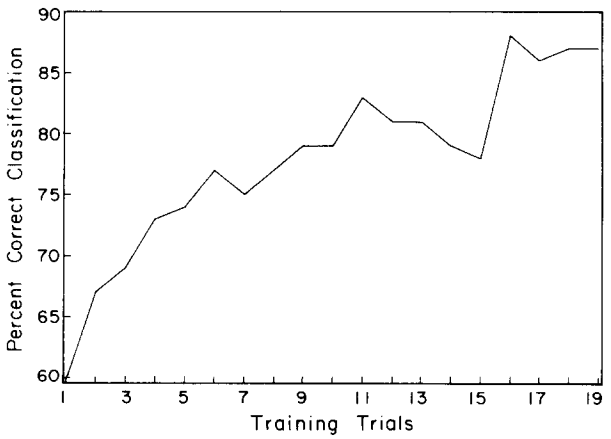

FIGURE 15. The training curve of a human listener trained to distinguish the two sonar targets on the basis of the same set of returns used to train the networks in the aspect-angle dependent series of experiments. Each trial consisted of 104 returns presented in a random sequence.

It was also found that, in addition to the general aspect-angle independent classification strategy, an aspect-angle dependent strategy was adopted in order to correctly classify less frequent signals that did not conform to the general model of a cylinder or rock return. This was accomplished by using a small number of weights to encode specific spectral peaks and nulls.

Although it is attractive to think of a hidden unit as a feature extractor, this may not be the best way to characterize a hidden unit's coding strategy. As we demonstrated, the hidden unit is capable of encoding multiple features and even multiple strategies simultaneously. This kind of pattern coding makes efficient use of the capacity of each hidden unit and is more suggestive of a model-based approach rather than simple feature extraction. The network is able to internally encode pattern variations that do not decompose simply into a set of feature dimensions. An important step toward understanding neural networks will be the development of an appropriate formalism for describing this coding strategy.

Comparison with Human Performance

The performance of human listeners on the same set of data was comparable to the network classifier.

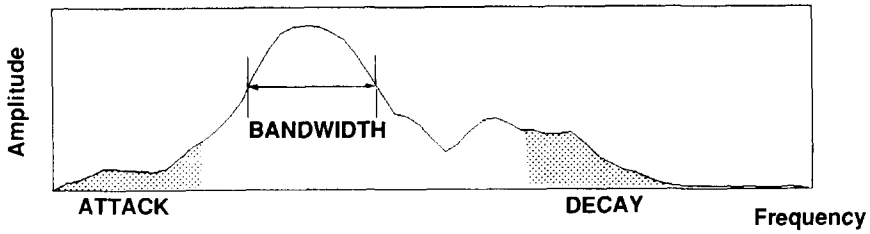

FIGURE 16. The spectral envelope of a typical sonar return. The signal features which correlated well with human perceptual cues are indicated.

Although it is difficult to make a direct comparison, due to the use of different signal representations, the high correlation between human perceptual ratings and network's hidden unit responses to the same set of signals suggests that a similar internal representation may underlie their comparable performance.

Although this is a limited study in many respects, the results suggest that network classifiers should provide a viable alternative to existing machine-based techniques. The performance of the network classifiers is better than a nearest neighbor classifier and less expensive in terms of storage and computation. In addition, the networks are able to achieve near optimal performance without requiring a priori knowledge or assumptions about the underlying statistical structure of the signals to be classified. Finally and perhaps most importantly, the network's performance and classification strategy appear to be comparable to that of humans.

REFERENCES

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans on Inform. Theory*, **IT-13,** 21–27.

Gorman, R. P., & Sawatari, T. (1987). Automatic sonar target recognition based on human perceptual features. Submitted to *Journal of the Acoustic Society of America.*

Gorman, R. P., & Sejnowski, T. J. (1987). Learned classification of sonar targets using a massively-parallel network. Submitted to *IEEE Trans on Acoustics Speech Signal Processing.*

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika,* **32,** 241–253.

Lapedes, A., & Farber, R. (1987). Nonlinear signal processing using neural networks: Prediction and system modelling. Submitted to *Proceedings of IEEE.*

Lehky, S. R., & Sejnowski, T. J. (1987). Extracting curvatures of 3-D objects using a neural network. *Society for Neuroscience Abstracts,* **13,** 1451.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE Acoustics Speech Signal Processing Magazine,* **4,** 4–22.

Rosenberg, C. R., & Sejnowski, T. J. (1987). Parallel networks that learn to pronounce English text. *Complex Systems,* **1,** 145–168.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Exploration in the microstructure of cognition* (pp. 318–362). Cambridge: MIT Press.

Sejnowski, T. J., Kienker, P. K., & Hinton, G. H. (1986). Learning symmetry groups with hidden units: Beyond the perceptron. *Physica,* **22D,** 260–275.

Toussaint, G. T. (1974). Bibliography on estimation of misclassification. *IEEE Trans on Inform. Theory,* **IT-20,** 472–479.

Watrous, R. L., & Shastri, L. (1987). Learning acoustic features from speech data using connectionist networks. *Proceedings of the Ninth Annual Conference of The Cognitive Science Society,* **9,** 518–530.

Widrow, G., & Hoff, M. E. (1960). *Adaptive switching circuits.* Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4 pp. 96–104.