In Proceedings of SPIE99 (vol. 3710), Florida, April, 1999

Feature extraction and fusion of wide-band backscattered signals

Nathan Intrator
Institute for Brain and Neural Systems
Brown University
Providence, RI 02912

Quyen Q. Huynh
Institute for Brain and Neural Systems and
Coastal Systems Station
Naval Surface Warfare Center
Panama City, FL 32407-7001

Gerald J. Dobeck
Coastal Systems Station
Naval Surface Warfare Center
Panama City, FL 32407-7001

April, 1999

Abstract

Good discrimination results have been obtained with an active backscatter data set of mine-like objects [1], where the task was to distinguish between man-made and non-man-made objects. In this work we introduce a novel method for constructing best basis for discrimination from wavelet packets, and demonstrate the superiority of multiple ensembles of predictors. We achieve far better discrimination results using a wide band FM sweep of 80kHz compared with the earlier work that used a 40kHz FM sweep. We further show improved results by combing several discriminative methods with several wavelet packet representations.

1 Introduction

Discrimination problems differ in nature from compression tasks. While in compression, it is the mean squared error that is often used to measure the quality of the scheme, classification requires a different measure which often is not related to the former. The discrimination power of a certain basis or a set of basis functions is not necessarily connected to the quality of compression associated with this set. Furthermore, the degree of relevance of the orthonormality constraint to the quality of the discrimination is questionable. For example, linear discriminant analysis [2] searches for linear projections which maximize the between-class variance divided by the sum of within-class variance. Such projections do not necessarily coincide with the principal components of the data which are the directions that optimize MSE compression.

A successful approach to discrimination is based on an appropriate preprocessing to create an efficient signal representation, which then leads to an efficient dimensionality reduction. The next step is again some combination of feature extraction and classification. In this paper we discuss methods for extracting features from wavelet representations for the purpose of discrimination between classes of signals. We propose a novel method for constructing best basis for discrimination between signals and then discuss performance improvement via ensemble averaging of collections

of experts. Results on the narrow- and wide-band backscatter sonar data are presented. briefly review several methods for finding data representation via optimal decomposition of wavelet basis functions and discuss their compression properties. We then discuss some signal decomposition methods for the purpose of discrimination. This is followed by a brief discussion on a combination of feature extraction and classification scheme and with discrimination results on two acoustic data-sets.

2 Optimal basis decomposition for discrimination

There has been a lot of work on optimal wavelet basis for compression stemming from the work of Coifman and Wickerhauser [3] which have presented a simple and fast algorithm for finding the local best basis (BB) in a wavelet packet (WP) library basis functions. Their algorithm which is based on comparing the entropy of different wavelet packet cells is very fast and thus allows for a quick comparison between an exponential number of possible bases.

Search in non-orthogonal bases is done using an over-complete dictionary of basis functions [4, 5, 6]. However, there has been much less work on optimal bases for signal discrimination. Below, we briefly review this work including our proposed algorithm.

2.1 Local discriminant bases

The local discriminant base (LDB) [7, 8] creates a time-frequency dictionary such as WP or local trigonometric functions (CP), from which signal energies for each basis coordinates are accumulated for each signal class separately. Then, a complete orthonormal basis is formed using a distance measure between the distributions of those energies from each class.

The original algorithm [7] attempted to extract best basis from the energies (squared values) of the WP, which is the direct approach to finding a best basis for a class of patterns [3]. Unfortunately, when the distance measure is applied to these energy coefficients, or more generally to the distribution of the energies, then the interpretation of the new basis is not clear anymore and the optimality properties are not so apparent. Moreover, noticing that the energies may not be so indicative for discrimination, Saito and Coifman [8] have suggested to use a different non-linear function of the basis function of the coefficients (instead of a just square values) so as to alleviate this problem.

2.2 Discriminant pursuit

Buckheit and Donoho [9] have introduced the discriminant pursuit (DP) algorithm. It follows the approach of basis pursuit, in the sense that it is not constrained by seeking only orthogonal discriminant basis functions, but can search in the over-complete WP or CP dictionary. The discrimination power of each basis function is measured by:

$$D_i(X,Y) = \frac{|E_X[wp_i(x)] - E_Y[wp_i(y)]|}{\mathrm{STD}_X(wp_i(x)) + \mathrm{STD}_Y(wp_i(y))}, \tag{1}$$

which is a 1-dimensional form of Fisher discriminant analysis criterion [2]. It is our experience that often, the additional flexibility of using an over-complete set instead of a basis, often leads to inferior results. This happens when the the number of training patterns is small relative to the signal dimensionality, and thus, there may be many wavelet basis functions whose energy is composed of very few training patterns. Such elements may be chosen for discrimination, although their overall contribution to the signal in miniscule. If the WP representation is sparse, then for every basis

function there are very few patterns which contribute to its value, thus the variability is large and outliers are more likely to cause trouble. There is another problem associated with this approach; Since the wavelet packet transformation is linear, it follows that $E_X[wp_i(x)] = wp_i(E_X[x])$. Thus, if the mean of each signal set is zero, there is no discrimination power in the means. A simple example is the discrimination between two signals of the form: $\sin(\omega t + u)$ and $\sin(2\omega t + u)$, where $u \sim U[0, 2\pi]$.

2.3 Quadratic Discrimination

From a given set of observations of one class, $\{X_i\}_{i=1}^n$ and a set of observations of another class, $\{Y_j\}_{j=1}^m$ it is possible to create a set of $n * m$ observations of the form

$$\{Z_{ij} = (X_i - Y_j)\}_{i,j=1}^{n,m}.$$

A low dimensional representation (LDR) for Z should have a good discriminating power between X and Y. Thus for example, one can consider the first few coefficients of a best basis for Z as such a LDR. The problem with this approach is that it appears that the number of calculations required to create a best basis for Z is $O(n * m)$, which may be prohibitively large. However, this calculation can actually be done in $O(max(n, m))$ calculations as is seen below.

Let $W(Z)$ denote the wavelet packet transformation of Z, then $W^2(Z)$ corresponds to the wavelet packet transformation of Z^2, where W^2 means that every wavelet packet basis is squared. For constructing a best basis of Z we need to calculate

$$\sum_{i=1}^{n}\sum_{j=1}^{m} W^2(Z_{ij}).$$

Consider a specific wavelet packet basis $wp_r(X_i - Y_j)$; Due to linearity of the wavelet packet operation we get

$$wp_r(X_i - Y_j) = wp_r(X_i) - wp_r(Y_j).$$

Thus,

$$\begin{aligned} wp_r^2(X_i - Y_j) &= [wp_r(X_i) - wp_r(Y_j)]^2 \\ &= [wp_r^2(X_i) + wp_r^2(Y_j) - 2wp_r(X_i)wp_r(Y_j)]. \end{aligned}$$

The mean contribution of all the signals to this wavelet packet entry is given by

$$\begin{aligned} \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} wp_r^2(X_i - Y_j) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[wp_r^2(X_i) + wp_r^2(Y_j) - 2wp_r(X_i)wp_r(Y_j)] \\ &= \frac{1}{n}\sum_i wp_r^2(X_i) + \frac{1}{m}\sum_j wp_r^2(Y_j) - 2[\frac{1}{n}\sum_i wp_r(X_i)][\frac{1}{m}\sum_j wp_r(Y_j)]. \end{aligned}$$

Thus the calculation of this best discriminating basis which we call quadratic discriminating basis (QDB) is as efficient as the calculation of the regular best basis, only that this one is specifically geared towards discrimination.

3 The Variance-Bias Dilemma

The motivation of our approach follows from a key observation regarding the bias/variance decomposition, namely the fact that ensemble averaging does not affect the bias portion of the error, but reduces the variance, when the estimators on which averaging is done are independent.

The classification problem is to estimate a function $f_{\mathcal{D}}(x)$ of observed data characteristics x, for predicting a class label y, based on a given training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_L, y_L)\}$, using some measure of the estimation error on $\mathcal{D}$. A good estimator will perform well not only on the training set, but also on new *validation* sets which were not used during estimation.

Evaluation of the performance of the estimator is commonly done via the mean squared error distance (MSE) by taking the expectation with respect to the (unknown) probability distribution P of y:

$$E[(y - f_{\mathcal{D}}(x))^2 | x, \mathcal{D}].$$

This can be decomposed into

$$E[(y - f_{\mathcal{D}}(x))^2 | x, \mathcal{D}] = E[(y - E[y|x])^2 | x, \mathcal{D}] + E[(f_{\mathcal{D}}(x) - E[y|x])^2].$$

The first term typically depends on neither the training data $\mathcal{D}$ nor the estimator $f_{\mathcal{D}}(x)$, it measures the amount of noise or variability of y given x. Hence f can be evaluated using

$$E[(f_{\mathcal{D}}(x) - E[y|x])^2].$$

The empirical mean squared error of f is given by

$$E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E[y|x])^2],$$

where $E_{\mathcal{D}}$ represents expectation with respect to all possible training sets $\mathcal{D}$ of fixed size.

To investigate further the MSE performance we decompose the error into bias and variance components [10] to obtain

$$E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E[y|x])^2] = (E_{\mathcal{D}}[f_{\mathcal{D}}(x)] - E[y|x])^2 + E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E_{\mathcal{D}}[f_{\mathcal{D}}(x)])^2]. \quad (2)$$

The first term on the right-hand side is called the bias term (strictly, the squared bias) of the estimator and the second term is called the variance term. When training on a fixed training set $\mathcal{D}$, reducing the bias with respect to this set may increase the variance of the estimator and contribute to poor generalisation performance. This is known as the trade-off between variance and bias. Typically, variance is reduced by smoothing, but this may introduce bias since, for example, it may blur sharp peaks. Bias is reduced by incorporating prior knowledge. When prior knowledge is used also for smoothing, it is likely to reduce the overall MSE of the estimator.

When training neural networks such as multilayer perceptrons, the variance arises from two terms. The first term comes from inherent data randomness and the second term arises from the non-identifiability of the model, in that, for a given training dataset, there may be several local minima of the error surface.

Consider the ensemble average $\bar{f}$ of Q predictors, which in our case can be thought of as neural networks with different random initial weights which are trained on data with added Gaussian noise:

$$\bar{f}(x) = \frac{1}{Q} \sum_{i=1}^{Q} f_i(x).$$

These predictors are identically distributed, and thus the variance contribution to equation (2) becomes

$$\begin{aligned} E[(\bar{f} - E[\bar{f}])^2] &= E[(\frac{1}{Q}\sum f_i - E[\frac{1}{Q}\sum f_i])^2] \\ &= E[(\frac{1}{Q}\sum f_i)^2] + \left(E[\frac{1}{Q}\sum f_i]\right)^2 - 2E[\frac{1}{Q}\sum f_i E[\frac{1}{Q}\sum f_i]] \end{aligned}$$

$$= \quad E[(\frac{1}{Q}\sum f_i)^2] - \Big(E[\frac{1}{Q}\sum f_i]\Big)^2; \tag{3}$$

we omit mention of x and $\mathcal{D}$ for clarity. The first term in (3) can be rewritten as

$$E[(\frac{1}{Q}\sum f_i)^2] = \frac{1}{Q^2}\sum E[f_i^2] + \frac{2}{Q^2}\sum_{i<j} E[f_i f_j],$$

and the second term gives

$$\Big(E[\frac{1}{Q}\sum f_i]\Big)^2 = \frac{1}{Q^2}\sum\Big(E[f_i^2]\Big)^2 + \frac{2}{Q^2}\sum_{i<j} E[f_i]E[f_j].$$

Plugging these equalities into (3) gives

$$E[(\bar{f} - E[\bar{f}])^2] \quad = \quad \frac{1}{Q^2}\sum\{E[f_i^2] - \Big(E[f_i]\Big)^2\} + \frac{2}{Q^2}\sum_{i<j}\{E[f_i f_j] - E[f_i]E[f_j]\}. \tag{4}$$

Set

$$\gamma = \mathrm{Var}(f_i) + (Q-1)\mathrm{max}_{i,j}(E[f_i f_j] - E[f_i]E[f_j]).$$

It follows that[1]

$$\frac{1}{Q}\mathrm{Var}(f_i) \leq \mathrm{Var}(\bar{f}) \leq \frac{1}{Q}\gamma \leq \max_i \mathrm{Var}(f_i). \tag{5}$$

This analysis suggests a simple extrapolation to large values of Q by giving an upper bound of $1/Q\gamma$ to the variance behaviour under large network-ensembles from small-size ensembles [11]. Note that

$$E[f_i f_j] - E[f_i]E[f_j] = E\Big(\{f_i - E[f_i]\}\{f_j - E[f_j]\}\Big).$$

Thus, the notion of independence can be understood as independence of the deviations of each predictor from the expected value of the predictor, which can be replaced, because of linearity, by

$$E\Big(\{f_i - E[\bar{f}]\}\{f_j - E[\bar{f}]\}\Big),$$

and is thus interpreted as an indication of the prediction variation around a common mean.

4 Ensemble fusion of experts

A general framework for combining multiple estimators is "Stacked Generalization" [12], where each estimator is trained with different sub-set of the data and the optimal combination is estimated using cross validation methods. this method has been compared with other ensemble training methods in [13].

We have introduced an *Integrated Classification Machine (ICM)* that is constructed of a hierarchy of classifiers [14] (Figure 1). The smallest building block is a conventional Feed-Forward Multi Layer Perceptron with Sigmoidal outputs. All networks are trained to predict the class label of the signal.

As was discussed in Section 3, reduction in the variance portion of the error is achieved when the errors made by different predictors are independent. This has been achieved for example by injecting

[1] We use the fact that $ab \leq \frac{a^2+b^2}{2}$, thus, $E[f_i f_j] - E[f_i]E[f_j] = E\Big(\{f_i - E[f_i]\}\{f_j - E[f_j]\}\Big) \leq \max_i \mathrm{Var}(f_i)$.

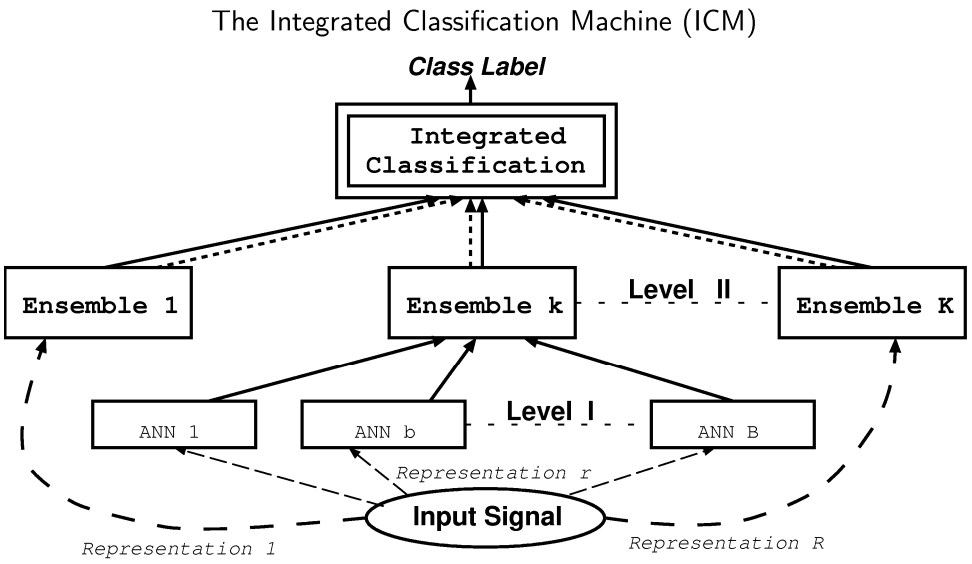

Figure 1: Several representations of the waveform are fed into different Ensembles, then integrated to produce the final classification. (In level II, the regular arrows are the Ensembles' prediction values and the dashed arrows are the attached confidence values).

noise into the inputs in [15]. Here we achieve some independence by using different mother wavelets, by extracting different basis representation and by feature extraction using different discrimination methods. Thus, the inputs to the different predictors in the *ICM* may have different dimensionalities according to the respective input representation used. Under a neural network setup, the hidden layers (of the different predictors) contain various numbers of sigmoidal units and the output layer contains two sigmoidal units. Further details are given in [14]. These include exact architectures, the construction of prediction ensemble, the usage of sets of ensembles to determine the confidence of the set. Section **??** demonstrates the effect of different mother wavelets and different feature extraction methods on the wide-band classification results.

5 Backscattered sonar data discrimination

This application involves an active backscatter data set of mine-like objects. The data was collected at the Naval Surface Warfare Center (NSWC) by Gerald Dobeck. The task was to distinguish between man-made and non-man-made objects. There were six objects in the data: metal cylinder, cone-shaped plastic object, water-filled barrel, limestone rock, granite rock, and a water-logged wooden log. The data-set contained seven different frequency bands. We report here results for two bands: a narrow band (20-60kHz) and a wide band (30-110kHz). The targets were suspended in a large water tank, while cylindrical objects were suspended horizontally. Measurements were collected in 5 degree increments on a rotating target around a vertical axis. Every second measurement was used for testing, thus the train and test data were interleaved and both included measurements at 10 degree increments. Synthesized reverberation at different SNR levels was added to the raw acoustic backscatter signals. The reverberation was synthesized by convolving white noise with the transmit pulse. To simulate a detection process, the noisy signals were deconvolved with the transmit pulse. The peak of the deconvolved signal was normalized to unity, and a 0.512 milli-second segment centered at the peak, was used for data analysis. Due to the noise simulation, there were

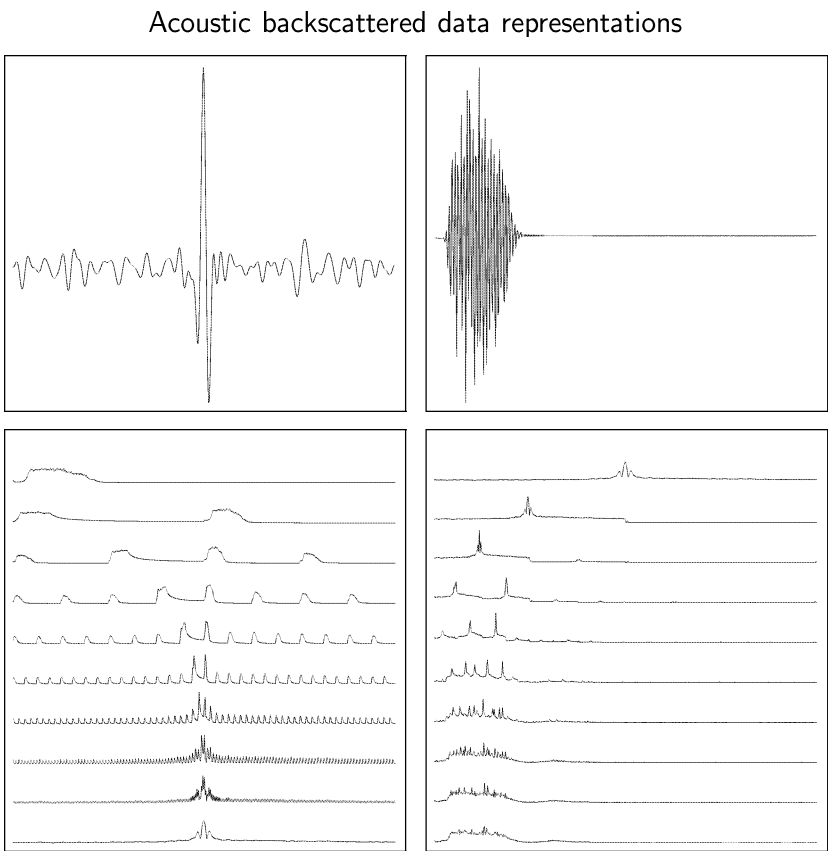

Figure 2: A typical signal from the data-set. Top left: Raw signal, right: Best Basis from Local Cosine. Bottom left: Local Cosine packet, right Wavelet Packet (Symmlet-6).

2160 patterns in the training and in the testing sets. We report here results with signal to noise ratio of 20db.

The signal/noise ratio was between 12-20db, where the noise was an accurate modeling of the bottom reverberation. It appears that there is very little sensitivity to the noise level between these levels. Further details about the data as well as other classification results are given elsewhere in this volume [16]. Figure 2 depicts various signal representations of the backscattered data. Next to the raw signal at the top, a best-basis representation based on Coifman's algorithm [3] is shown. This basis was extracted from a local cosine packet that is presented at the bottom left panel. A wavelet packet representation using a Symmlet-6 mother wavelet is shown on the bottom right.

The results of the narrow-band backscattered data are in Table 1. It appears that a local cosine representation is more appropriate for this acoustic signal and several discriminant basis selections yield similar results. These results reflect an improvement of around 3% which is due to BCM constraints[2] and ensemble averaging over noise injected inputs [18, 15, for implementation details]. The results are generated assuming that false positive and false negative errors have the same weight. Surprisingly, linear discrimination from wavelet packet representation as suggested by [9] did not yield good results in any of the wavelet representations. This may be due to the small size of the training set which may lead to a very sparse representation in which a small number of training patterns cause a nonzero value for some of the wavelet basis and appear to have a high

[2]BCM constraints bias the model towards finding low dimensional representations which are multi-modal and have high kurtosis [17].

Backscattered data classification

Narrow-band (40kHz)

	Local Cosine	Symmlet-6	Coiflet-3
BB	18.6(1.2)	33.1(1.4)	41.1(1.4)
LDB	18.7(0.8)	40.0(1.4)	39.9(1.6)
Wavelet	28.9(1.3)	42.7(1.8)	41.8(1.0)
QDB	18.7(1.5)	29.9(1.3)	30.3(0.8)

Table 1: Percent classification error rate of mine-like object discrimination using various discriminant basis representations from a Coiflet-3 and Symmlet-6 mother wavelets and a local cosine wavelet packet. The discrimination is performed from most discriminative 15 coefficients extracted by each method.

discrimination value.

It can be expected that combining information from several frequency bands will improve these results.

5.1 Wide-band results

Wide band back-scattered data has not been extensively analyzed as the narrow band sonar. Recently, the development of new materials and composites has made the development of broadband sonars possible, especially in the higher frequency range. The current broad-band data is composed of a 1 milli-second long FM chirp between 30kHz-110kHz. Preprocessing and general experimental conditions are similar to those of the narrow-band data.

We have achieved far better results on this frequency band, and thus provide more results to demonstrate the effect of different mother wavelets, different discrimination methods and different number of features. Table 2 provides results of the experiments with the wide-band data. The top panel indicates that vast dimensionality reduction is essential for improved performance as the discriminating information is confined in a much smaller dimensional space. The bottom panel presents results of 12 and 15 features. The Wavelet^1 features were the first 12 or 15 coefficients of the wavelet basis using symmlet-8 or coiflet-5 mother wavelets. Wavelet^2 features were the most discriminating features using Fisher linear discrimination [2] of each wavelet basis function separately. QDB^1 means that a quadratic discrimination best basis is used and the highest energy coefficients are chosen. In QDB^2, the discriminating coefficients, which are extracted from the same quadratic discrimination bases, are those that maximize the L^2 distance of the distributions[3] of the two classes (on the training set) In all cases the discrimination is done on each dimension separately.

The ensemble result in each row is an ensemble of the networks that were used in that row. Each entry in the table is a result of an ensemble of 7 feed-forward networks which differ on the initial conditions only. Thus each entry in the table corresponds to the second level in the integrated classification machine (Figure 1). The ensemble and super ensemble entries correspond to the top level of the ICM. The ensemble entry combines networks with the same number of input features, while the super ensemble combines networks of the 12 and 15 features together. The results indicate that ensemble averaging is very powerful, in improving and robustifying the results. It further shows that ensemble results can be improved even if networks that do not perform well as single predictors are included. In short, the results indicate that ensemble averaging may be the solution to the need

[3]We have found this measure to be more robust than the KL divergence between the distributions.

Backscattered data classification

Wide-band (80KHz)

	Symmlet-8 512 Coeffs	Symmlet-8 128 Coeffs	Symmlet-8 20 Coeffs
Wavelet^1	33.0	30.1	22.3
Wavelet^2	30.7	9.8	8.2

15 Coeffs	Symmlet-8	Coiflet-5	Ensemble
Wavelet^1	7.3	6.0	6.3
Wavelet^2	9.2	7.0	6.9
QDB^1	7.0	6.5	6.3
QDB^2	9.6	6.9	7.4
Ensemble	6.4	5.7	5.7
12 Coeffs			
Wavelet^1	8.6	6.9	7.8
Wavelet^2	15.5	14.0	14.4
QDB^1	7.3	8.3	7.4
QDB^2	11.9	7.5	8.0
Ensemble	6.9	6.3	6.6
Sup Ensemble	6.7	5.4	6.1

Table 2: Percent error classification. Numbers in brackets represent standard deviations. Wavelet^1 uses the first 15 wavelet coefficients of the signal while Wavelet^2 stands for a discrimination based on L^2 distance of the distributions of the two classes. QDB^1 means that first the quadratic discrimination best basis is found (see text for details) and then the first 12 or 15 coefficients of this basis are used. In Quadratic^2, the discriminating coefficients based on the L^2 distance of the distributions of the two classes are found from the same best basis.. In all cases the discrimination is done on each dimension separately.

to estimate various model parameters such as the optimal number of features, the optimal mother wavelet, the optimal discrimination method etc.

Acknowledgments

This work, including the collection of the data, was supported by the Defense Advanced Research Project Agency (DARPA-DSO) and the Office of Naval Research. Many discussions with various members of the Institute for Brain and Neural Systems at Brown University are gratefully acknowledged.

References

[1] N. Intrator, Q. Q. Huynh, and G. Dobeck, "Feature extraction from acoustic backscattered signals using wavelet dictionaries," in *Proceedings of SPIE97*, Florida, April 1997, vol. 3079, pp. 183–190, IEEE.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[3] R. R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Info. Theory*, vol. 38, no. 2, pp. 713–719, 1992.

[4] I. Daubechies, "Time-frequency localization operator: a geometric phase space approach," *IEEE Transactions on Information Theory*, vol. 34, pp. 605–612, 1988.

[5] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[6] S. S. Chen, D. L. Donoho, and M. Saundres, "Atomic decomposition by basis pursuit," Technical Report, Stanford University, February 1996.

[7] N. Saito and R. R. Coifman, "Local discriminant bases," in *Proc. SPIE 2303*, A. F. Laine and M. A. Unser, Eds., 1994, pp. 2–14.

[8] N. Saito and R. R. Coifman, "Improved local discriminant bases using empirical probability density estimation," in *Amer. Stat. Assoc. Proceeding on Statistical Computing*, 1996.

[9] J. Buckheit and D. L. Donoho, "Improved linear discrimination using time-frequency dictionaries," Technical Report, Stanford University, 1995.

[10] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias-variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1992.

[11] U. Naftaly, N. Intrator, and D. Horn, "Optimal ensemble averaging of neural networks," *Network*, vol. 8, no. 3, pp. 283–296, 1997.

[12] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[13] M. LeBlanc and R. Tibshirani, "Combining estimates in regression and classification," 1994, Preprint.

[14] Y. Shimshoni and N. Intrator, "Classifying seismic signals by integrating ensembles of neural networks," *IEEE-Signal Processing*, vol. 46, no. 5, pp. 1194–1201, May 1998, `ftp://cns.brown.edu/nin/papers/p.ps.Z`.

[15] Y. Raviv and N. Intrator, "Bootstrapping with noise: An effective regularization technique," *Connection Science, Special issue on Combining Estimators*, vol. 8, pp. 356–372, 1996.

[16] L. L. Burton and H Lai, "Active sonar target imaging and classification system," in *Proceedings SPIE*, 1997, To appear.

[17] N. Intrator and L. N Cooper, "Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions," *Neural Networks*, vol. 5, pp. 3–17, 1992.

[18] N. Intrator, "Feature extraction using an unsupervised neural network," *Neural Computation*, vol. 4, pp. 98–107, 1992.