To appear: Proceedings of The 1997 Canadian Workshop
on Information Theory. June, 1997, Toronto.

Wavelet Feature Extraction for Discrimination Tasks

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

Quyen Q. Huynh*†
Coastal Systems Station
Naval Surface Warfare Center
Panama City, FL 32407-7001

Yau Shu Wong‡
Department of Mathematical Sciences
University of Alberta
Edmonton, CANADA

B. H. K. Lee
Institute for Aerospace Research
National Research Counsil of Canada
Ottawa, CANADA

1 Introduction

Discrimination problems differ in nature from reconstruction tasks. While in reconstruction, 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 function is not necessarily connected to the quality of reconstruction 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 [1] 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 reconstruction.

There have been several approaches to searching for basis functions for discrimination; Coifman adopts the orthonormal basis approach and is actually searching for a basis that best reconstructs the mean difference between two classes.

In this paper we briefly review several methods for finding optimal decomposition via basis functions and discuss their reconstruction properties. We then discuss some signal decomposition methods for the purpose of discrimination followed by discrimination results. The last two section describe a different application of wavelet representation to model estimation.

*Supported by ONR.

†Also with the Applied Mathematics Division, Brown University, Providence RI.

‡Supported by a grant from the Natural Sciences and Engineering Research Council, Canada.

2 Optimal basis function decomposition for reconstruction

2.1 Entropy based algorithms

Coifman and Wickerhauser [2] presented a simple and fast algorithm for finding the local best basis (BB) in a wavelet packet (WP) library basis functions. The search is very simple and fast due to the orthogonality condition between the basis functions at each level and the inclusion properties of basis functions between different levels. Choosing between different possible bases is done via the entropy of the coefficients, namely the speed of decay in coefficient values, which indicated the degree of compression of the representation.

2.2 Basis pursuit

Unlike the search in orthogonal bases as done in the best basis method, one can search in an overcomplete dictionary of basis functions. This has been proposed by Daubechies and termed the Method of Frames [3]. Among different representations for the same signal, one searches for a representation whose vector of coefficients has the smallest l^2 norm. This approach leads to a quadratic optimization problem that is solved via a system of linear equations. Recently Chen et al. [4] presented a Basis Pursuit method (BP) that is very similar to the methods of frames; It decomposes a signal using dictionary elements so that the coefficients have the smallest l^1 norm among all such decompositions. This optimization can be performed by recent linear programming techniques [5]. Chen et al. demonstrate that for certain signals, the convergence of a basis pursuit algorithm is faster than that of a best basis representation.

2.3 Matching pursuit

The Matching Pursuit algorithm [6] is an iterative algorithm, which does not explicitly seek any overall goal, but merely applies a simple rule repeatedly. It is a forward

model selection that adds at each step the single most correlated new atom among all those not included yet in the model. The algorithm is very powerful for orthogonal basis selection, but may fail for non-orthogonal dictionaries.

3 Optimal basis decomposition for discrimination

3.1 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 [2]. 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. However, this approach takes us even further away from interpretation and optimality of the best basis approach.

3.2 Discriminant pursuit

Buckheit and Donoho [9] have introduced the discriminant pursuit (DP) algorithm which 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 overcomplete WP or CP dictionary. The discrimination power of each basis function is measured by:

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

which is a 1-dimensional form of Fisher discriminant analysis criterion [1]. It is our experience that often, the additional flexibility leads to inferior results. This happens when the dimensionality is high and the number of training patterns is relatively small. 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 $u \sim U[0, 2\pi]$.

3.3 Applicability of wavelet representation to discrimination

The choice of optimal bases for discrimination may not be so practical for the reasons described above, namely, due to the small number of training patterns and large number of WP or CP coefficients that have to be estimated, resulting in overfitting to the training set. One way suggested by Buckheit and Donoho [9] is to "remove" the noise from the signal using de-noising. We present here a simple alternative: no basis optimization for the set of signals, rather a usage of a *good* general basis, namely, wavelets. We show (Table 1) that classification results and feature extraction from this basis may be superior to an attempt to optimize the basis for the class discrimination. This is due to the nonlinear separability in wavelet space which is not well captured by linear separation methods.

4 Non-linear feature extraction from wavelet representation

In this section we briefly discuss an unsupervised learning algorithm which searches for multi-modality in the projection space. Exploratory projection pursuit theory [10, 11] tells us that search for structure in input space can be approached by a search for deviation from normal distribution of the projected space. Furthermore, when input space is clustered, a search for deviation from normality can take the form of search for multi-modality, since when clustered data is projected in a direction that separates at least two clusters, it generates multi-modal projected distributions.

It has been recently shown that a variant of the Bienenstock, Cooper and Munro neuron (BCM) [12] performs exploratory projection pursuit using a projection index that measures multi-modality [13]. This neuron allows modeling and theoretical analysis of various visual deprivation experiments and is in agreement with the vast experimental results on visual cortical plasticity. A network implementation which can find several projections in parallel while retaining its computational efficiency, was found to be applicable for extracting features from very high dimensional vector spaces [14, 13]. This method is applied to feature extraction in a problem discussed in the next section.

4.1 Application to acoustic signal discrimination

The types of signals explored in this study are the marine mammal sounds of porpoise and sperm whale which were recorded at a sampling rate of 25 kHz at various locations such as the Gulf of Maine, the Mediterranean and the Caribbean sea. We consider large data files where the signal consist intermittently of mammal sounds and background noise. Each of these files contains whale or porpoise sounds, but not both. Several data sets of length

Feature Extraction From Time-Frequency Dictionary

	Porpoise	Whale
LDA on wavelet packet	94	33
LDB on wavelet packet	98	51
Highest energ. from wavelets	72	47
BCM extraction from wavelets	99	76
BCM applied on raw signals	32	95

Table 1: Results of linear and nonlinear feature extraction from wavelets and wavelet packet representation of Porpoise/Whale acoustic signals. LDA is the linear discriminant analysis of Buckheit and Donoho, LDB is the local discriminant basis of Saito and Coifman. BCM is a nonlinear feature extraction that searches for multi-modality (see text for details).

32768 samples corresponding approximately to 1.3 seconds, were extracted from these large files. These data sets which contained mammal sounds mixed with background noise, were used for training and testing.

Full discussion of the results appears in [15]. In this paper we only point out the fact that a choice of basis functions using a discrimination measure may not lead to best results and that optimizing (nonlinear) discrimination based on linear combinations of basis functions from a fixed (wavelet) basis, may be more effective.

5 Coherent structure extraction

We follow here the algorithm proposed by Coifman and Wickerhauser [16] for de-noising a given signal f of length N so that various parameters of a physical system can be estimated accurately. The noise is peeled off iteratively by projecting the signal on a sequence of optimal bases. The following decomposition process is done iteratively based on the signal-plus-noise model: $f = c_1 + r_1$, where the coherent part is c_1 and the residue (the noisy part) is r_1. For the next step, the residue r_1 is considered as a new signal which is decomposed as $r_1 = c_2 + r_2$. If this decomposition is repeated k times, we sum all the coherent parts: $c = c_1 + c_2 + ... + c_k$, then f is rewritten as $f = c + r_k$. Recall that from a given mother wavelet, we can construct a library of orthonormal bases e.g. wavelet packet and cosine packet. Therefore, we have at our disposition a large collection of libraries of bases. If we choose a library of bases, we search for the best basis B_i for the signal f in this library. We reorder the coefficients $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_N$ in decreasing order, which correspond to the basis B_i (with $b's$ as the basis functions). Then we pick the top M ($< N$) coefficients $\alpha's$ where the rate of decay is steepest. In the first equation, c_1 represents the reconstructed portion (coherent part) of f, which is based on these M coefficients $\alpha's$:

$$c_1 = \sum_{1}^{M} \alpha_j b_j \qquad (2)$$

Then r_1 is the residual vector (incoherent):

$$r_1 = \sum_{M+1}^{N} \alpha_j b_j \qquad (3)$$

The next step is to consider r_1 as a new signal for which we repeat the decomposition into a coherent part and an incoherent part (noisy part). Again we choose the best basis which is different from the previous best basis. Also the new basis can be from a new library. At each step, after reordering the coefficients, it is important to pick the largest M coefficients $\alpha's$ where the largest rate of decay occurs. If r_k represents an incoherent (e.g. gaussian) signal, any basis B_i in the library will not compress it very well. In fact, this is the stopping criteria for the iterative procedure. The true coherent part is the sum of all the individual coherent parts which are extracted during the iteration process.

6 Application to flutter analysis

To determine the flutter boundary of an aircraft requires accurate measurements of frequencies and damping values of critical vibration modes as a function of the flight velocity. Numerous techniques have been reported for real-time flutter identification with varying degree of success [17, 18]. In recent years, it has become clear that advances in wavelet theory for signal processing and the use of artificial neural networks to model complex characteristics of nonlinear systems have an important and direct relevance to parameters extraction of flutter signals.

The main goal of this section is to present the development of using wavelet and artificial neural networks to predict the frequencies and dampings of a simulated flutter signal. Data from a typical flight test usually includes responses from more than one mode of vibrations and can be expressed as:

$$y(t) = \sum_{i=1}^{n} a_i e^{-\alpha_i t} \sin(\omega_i + \phi_i) \qquad (4)$$

where n denotes the number of modes, a_i, α_i, ω_i and ϕ_i represent the amplitude, damping, frequency and phase angles of the i'th component. Note that, $\omega = 2\pi f$, where f is the frequency. Consider for simplicity a simple model consisting of two modes only:

$$y(t) = a_1 e^{-\alpha_1 t} \sin(\omega_1 t + \phi_1) + a_2 e^{-\alpha_2 t} \sin(\omega_2 t + \phi_2). \qquad (5)$$

The two exponentially decaying sine waves model the decaying portion of the response signals from the sine dwell or sine sweep excitations of the aircraft. Given a time series of such signal, our task is to determine the values of frequency and damping of the signal. The procedure under investigation is to apply artificial neural network used in conjunction with wavelet packet.

First, by using wavelet packets, we separate the two-mode signal into two one mode signals each containing one value of frequency and one value of damping coefficient. In fact, projected on the best basis, the two-mode signal exhibits clearly two distinct patterns on the phase plane, which are well separated both in time and frequency. By picking the highest coefficients corresponding to each pattern, the reconstruction gives each of the desired exponentially decaying sine wave.

An application of the techniques presented in the previous section is to embed the signal in noise. Again, we could peel off noise and retain only the exponentially decaying sine waves.

Next, we use a two-layer (one hidden layer and one output layer) artificial neural network with feed-forward connections. A sigmoid transfer function is employed, and the conjugate gradient algorithm is used to minimize the performance index which represents the square of the errors. To reduce the complexity of an artificial neural network, a wavelet transform is applied to the original signal, and we select only m largest wavelet coefficients as input to our neural network. The value of m is usually small, and it is certainly much smaller than M, the number of the data points for the original signal. Consequently, the number of inputs and hidden units is small. This results in a more efficient and robust network model.

In our computational experiments, the input layer is of 20–30 dimensions and the hidden units layer contains 15–20 neurons. Using 200 data sets for the training, the network is then tested on 20 testing data sets. The relative errors for the predicted damping coefficients are within 5%, and the relative errors for the predicted frequency values are within 3%.

The most attractive feature proposed here is that when dealing with real world problem, i.e., multi-mode signals, our method can be naturally and effectively implemented in parallel. Since a multi-mode signal is first decomposed into single-mode components using a wavelet routine, our artificial neural network can then be employed to extract the frequency and damping associated with each one-mode signal. The process can certainly be done in parallel. In future studies, we plan to use our algorithm for multi-mode signal problems and to compare the efficiency and accuracy by using artificial neural network directly to predict the parameter values of a multi-mode signal data.

References

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

[2] 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.

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

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

[5] P. E. Gill, W. Murrary, and M. H. Wright, *Numerical Linear Algebra and Optimization*. Addison Wesley, Redwood City, CA, 1991.

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

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

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

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

[10] P. J. Huber, "Projection pursuit. (with discussion)," *The Annals of Statistics*, vol. 13, pp. 435–475, 1985.

[11] J. H. Friedman, "Exploratory projection pursuit," *Journal of the American Statistical Association*, vol. 82, pp. 249–266, 1987.

[12] E. L. Bienenstock, L. N Cooper, and P. W. Munro, "Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex," *Journal Neuroscience*, vol. 2, pp. 32–48, 1982.

[13] 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.

[14] N. Intrator and J. I. Gold, "Three-dimensional object recognition of gray level images: The usefulness of distinguishing features," *Neural Computation*, vol. 5, pp. 61–74, 1993.

[15] Q. Huynh, L. N. Cooper, N. Intrator, and H. Shouval, "Classification of underwater mammals using feature extraction based on time-frequency analysis and bcm theory," 1996. To appear IEEE-SP, Special Issue on NN Applications.

[16] R. Coifman and M. V. Wickerhauser, "Wavelets and adapted waveform analysis. a toolkit for signal processing and numerical analysis," in *Different perspectives on wavelets, I* (I. Daubechies, ed.), pp. 119–145, Providence: American Mathematical Society, 1993.

[17] R. Walker and N. Gupta, "Real time flutter analysis," Technical Report, NASA CR-170412, 1984.

[18] B. H. K. Lee and F. Laichai, "Development of post-flight and real-time flutter analysis methodologies," in *Proceedings Forum International Aeroelasticite et Dynamique de Structures, Strasbourg*, pp. 703–719, 1993.