APPLIED AND COMPUTATIONAL HARMONIC ANALYSIS 2, 265–278 (1995)

A Wavelet Packet Approach to Transient Signal Classification[1]

RACHEL E. LEARNED AND ALAN S. WILLSKY

Massachusetts Institute of Technology

Communicated by Iain M. Johnstone

Received September 29, 1993; revised February 2, 1995

Time–frequency transforms, including wavelet and wavelet packet transforms, are generally acknowledged to be useful for studying non-stationary phenomena and, in particular, have been shown or claimed to be of value in the detection and characterization of transient signals. In many applications time–frequency transforms are simply employed as a visual aid to be used for signal display. Although there have been several studies reported in the literature, there is still considerable work to be done investigating the utility of wavelet and wavelet packet time–frequency transforms for *automatic* transient signal classification. This paper contributes to this ongoing investigation through the development of a non-parametric wavelet packet feature extraction *procedure* which identifies features to be used for the classification of transient signals for which explicit signal models are not available or appropriate. Recent literature in this area is devoted to truly ad-hoc, high-dimensional, non-parametric types of classification in which one or more time–frequency transform forms the base from which a large number of features are determined by trial and error. In contrast, the wavelet-packet-based procedure presented in this paper was formulated to *systematically adapt* to any data dictionary within which several classes must be distinguished. This method is aimed at focusing the information in the data set to find the smallest number of features for robust, reliable classification. The promise of our method is illustrated by performing our procedure on a set of biologically generated underwater acoustic signals. For this example the wavelet-packet-based features obtained by our method yield excellent classification results when used as input for a neural network and a nearest neighbor rule.

1. INTRODUCTION

Signals possessing non-stationary information are not suited for detection and classification by traditional Fourier methods. An alternate means of analysis needs to be employed so that valuable time–frequency information is not lost. The wavelet packet transform (WPT) is one such time–frequency analysis tool. This paper examines the feasibility of using the WPT in automatic transient signal classification through the development of a simple and systematic non-coherent feature extraction procedure which is tested on biologically generated underwater acoustic transient signals in ocean noise.

The classification of transient signals is both an interesting and important problem. Specifically, the ability to classify underwater acoustic signals is of great importance to the Navy. Today, detection and classification, tailored for stationary signals, is done by Naval personnel who listen to incoming signals while viewing computer generated displays which plot time vs angle-of-arrival and time vs frequency. The signal of interest is monitored and the primary frequencies contained in the signal are noted. An initial guess as to the source is made. In efforts to confirm or contradict the guess, the Naval officer will, perhaps repeatedly, consult tables which contain frequency information on a large range of possible signals.

Transient signals, lasting only a fraction of a second, are of particular concern because they will typically appear as broadband energy on the frequency display. Thus, the Naval officer cannot rely on any visual displays for assistance in the classification process. At present the human observer must be able to detect and classify transient signals by only listening for them. These brief signals may be missed by the listener. An automatic method of classification for transient signals would greatly aid in the detection/classification process.

A frequency display which uses standard spectral analysis methods is useful for stationary signal classification, but transient signals are not well matched to these methods. In particular, Fourier-based methods are ideally suited to the extraction of narrow band signals whose durations exceed or are at least on the order of the Fourier analysis window length. That is, Fourier analysis, particularly the short-time

[1] The work of the first author was supported by the Charles Stark Draper Laboratory under a research fellowship. The work of the second author was supported in part by the Draper Laboratory IR&D Program under agreement DL-H-467133, by the Air Force Office of Scientific Research under Grant AFSOR-92-J-0002, and by the Army Research Office under Grant DAAL03-92-G-0115.

1063-5203/95 $12.00

Fourier transform (STFT), does an excellent job of focusing the information for sources of this type, thus, providing features (spectral amplitudes) perfectly suited to detection and discrimination. For *transient* signals, the STFT with its non-varying window is not readily adaptable for capturing signal-specific characteristics. The STFT does allow for some temporal as well as frequency resolution, but it is not well suited for the analysis of many transient signals and, in particular, to the generation of features for detection and discrimination. This is discussed in greater detail in later sections.

The wavelet packet transform (WPT) offers a great deal of freedom in dealing with different types of transient signals. Indeed, the development of the wavelet transform (WT) [4, 10, 11, 15, 16] and wavelet packets [3, 23] has sparked considerable activity in signal representation and in transient and non-stationary signal analysis [1, 20, 21].

This paper is particularly interested in the research that has dealt with *automatic* detection and classification of transients. These works can roughly be grouped into two categories. One group of methods has focused on problems for which the classes of transients to be detected are well characterized by prior parametric models that identify the distinguishing characteristics of each class. Such methods generally operate based on coherent processing, i.e., on using wavelets as the basis for detection procedures that resemble matched filtering [7–9, 20]. In particular, Friedlander and Porat [7] find the optimal detector via the generalized likelihood ratio test for three linear time–frequency transforms of the received signal which is characterized by a signal model and a mismatch error in additive white Gaussian noise. They examine the performance of their detector with the STFT, the Gabor transform, and the WT. Frisch and Messer [8, 9] also formulate a detector by using the generalized likelihood ratio test for the WT coefficients of the received signal model. They restrict their signal model to an unknown transient with known relative bandwidth and time-bandwidth product. This assumption greatly reduces the complexity of the detector.

The second set of techniques, into which this research falls, deals with the detection and classification of transient signal classes that are not well-characterized in terms of prior models [2, 5, 12, 14, 17, 19]; consequently, somewhat different methods of detection and classification must be developed. In particular, recent work in the area of underwater acoustic transient classification using wavelet related concepts has been done by Nicolas, Lemer, and Legitimus [19] and Desai and Shazeer [5]. Both [19] and [5] employ wavelets as a means of generating features from various classes of underwater acoustic transients for input to a neural network. The authors of [19] use the energy in the wavelet decomposition of the transients along with features derived from autoregressive signal models and histograms of the data. The authors of [5] use the eight signals resulting from the third level of the wavelet packet decomposition (WPD), i.e., each transient signal is separated into eight components, one corresponding to each of eight equal bandwidth channels. The Fourier transform and curve length of these eight sequences are used as features. In the same spirit, much of the literature in the area of transient classification is devoted to the ad hoc development of features from one or more time–frequency transform. See [2, 12, 14, 17].

One characteristic common to the above efforts is that the transform used prior to feature construction is not adaptable to the signal characteristics. The WPT is the only time–frequency transform with a potential for adaptability.[2] In these works exploitation of class dependent frequency characteristics is suppressed by using a predetermined time–frequency transform or wavelet packet basis. In general, many researchers have come to the conclusion that no single time–frequency transform works best on all types of transients. Current methods for transient signal classification are developed by the trial and error approach to feature extraction resulting in procedures which are unique to each transient type on which they are performed and to each person that administers them. Moreover, it is typical for these ad hoc approaches to lead to the use of a large number of features (sometimes greater than the number of samples in the signal). This is a drawback in that pattern recognition algorithms work best with a reasonable number of inputs (features).

This paper describes an approach to signal classification based on the need for a non-parametric feature extraction algorithm that best adapts to sets of pre-classified data. As detailed in later sections, the WPT is well suited to a careful search for features. In contrast to the trial and error, high-dimensional, non-parametric types of classification which are truly ad-hoc approaches, the wavelet packet feature extraction technique presented in this paper is *systematic*. The feature extraction procedure does an orderly compilation of the information present in the data, *focusing* the information present in a set of signal samples into a *few* robust and well grounded statistics.[3] The promise of our method is illustrated by performing our procedure on a specific set of biological data. While the experimental results do represent tests on a substantial data set, the application results presented are preliminary. The purpose of our paper is to show the promise of the wavelet packet feature extraction procedure, offer motivation for further investigation of the utility of our method, and motivate research in using the

[2] Although, Desai and Shazeer did use the WPT, the choice of the basis was not considered as part of the feature selection process.

[3] In comparison, other non-parametric methods leave the task of sorting through the sets of signal samples (and often their Fourier and a variety of time–frequency transforms) to a neural network which, in general, is not to be reliable in cases where the number of inputs is large (this is due to the network settling in a local minima).

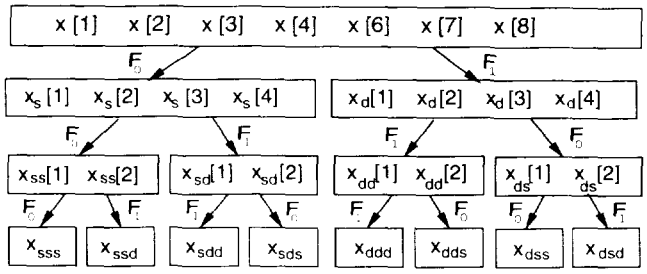

FIG. 1. The fully decomposed wavelet packet tree for a signal of length 8.

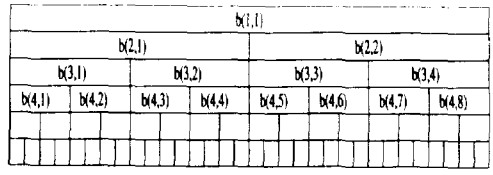

FIG. 2. The WPD tree with index label at each bin in the first four levels.

wavelet packet transform with systematic procedures for the development of a feature extraction algorithm that can be used on a broad class of transient signals.

This paper is organized as follows. Section 2 summarizes wavelet packet notation and establishes the energy mapping of the wavelet packet transform used in this paper. The wavelet packet feature extraction procedure is presented in Section 3. Section 4 illustrates our systematic method for determining wavelet-packet-based features by the formulation of a feature set for a set of biological transients. The Charles Stark Draper Laboratory and the Naval Undersea Warfare Center furnished an extensive collection of underwater acoustic signals in background noise which allowed for an empirical study of our feature extraction procedure on some typical occurrences of snapping shrimp and whale clicks. Using these features with a nearest neighbor rule and a neural network yielded 98% to 99% classification. Section 5 offers concluding remarks and a discussion of possible future work.

2. THE WAVELET PACKET TRANSFORM AND ITS ENERGY MAP

In this section we briefly review the structure of the discrete WPD that was developed by Coifman and Wickerhauser in [3]. We also introduce the notation and quantities to be used in the rest of this paper. WPD can be viewed as a natural extension of the WT providing a level by level transformation of a signal from the time domain to the frequency domain. The top level of a WPD is the time representation of the signal. As each level of the decomposition is calculated there is a decrease in temporal resolution and a corresponding increase in frequency resolution.

The wavelet decomposition of a discrete signal $x[n]$ may be calculated using a recursion of filter-decimation operations. Figure 1 shows a WPD tree for a signal of length eight.[4] The full WPD is displayed as a tree with a discrete sequence at every branch. F_0 and F_1 are the operators which perform the lowpass-downsampling and highpass-downsampling, respectively.[5] Each branch sequence is referred to as a *bin vector.* The decomposition may be continued down to the final level where there is only one element in each bin vector.

The bin locations within a tree will be represented by the notation b(l, c) where a bin is indexed by two parameters, level, l, and column, c. Figure 2 shows each bin of a WPD tree labeled with the appropriate bin position notation.

A few examples will illustrate the display used for the WPD of a signal and the time–frequency trade-off inherent in the WPD. Each bin vector of the WPD tree is displayed as a rectangular intensity plot at its appropriate position in the tree.

Figure 3 shows a signal comprised of two sinusoids and its full WPD. As the levels of the WPD tree are traversed, the information becomes more focused. The lowest level of the tree essentially agrees with the windowed discrete Fourier transform of the signal. Figure 4 shows a time and frequency localized signal corresponding exactly to one of the wavelet packet basis functions. Note the focusing of information at bin b(5, 6) of the tree. The information is less focused at the top and bottom of the tree, thus, the most compact or focused representation would be at b(5, 6) of the WPD tree.

Three points about these examples are worth noting. First, recall that the *wavelet* transform corresponds to a particular set of bins, namely those corresponding to successive low-pass/decimation (F_0) operations followed by a *single* highpass/decimation (F_1) operation. As pointed out in [3], only certain types of signals are well-focused in these bins. For example, the signal in Fig. 4 is focused at b(5, 6) which is *not* part of the ordinary wavelet decomposition. Second, as discussed in greater detail in Section 3.5, the STFT of a signal will not exhibit the same type focusing unless the most appropriate window size is used. The full WPT roughly corresponds to many STFT's using different window sizes. Third, the principle idea that we wish to exploit in finding useful features for transient detection and classification is precisely this focusing property, i.e. transients

[4] The display of the WPD trees throughout this paper is in accordance with the Paley ordering of the bins as described in [3].

[5] The Daubechies 14-point filter [4] is used for all the wavelet packet decompositions in this work. The choice of the wavelet used was somewhat arbitrary, guided primarily by a desire to have significant smoothness but reasonably short support.

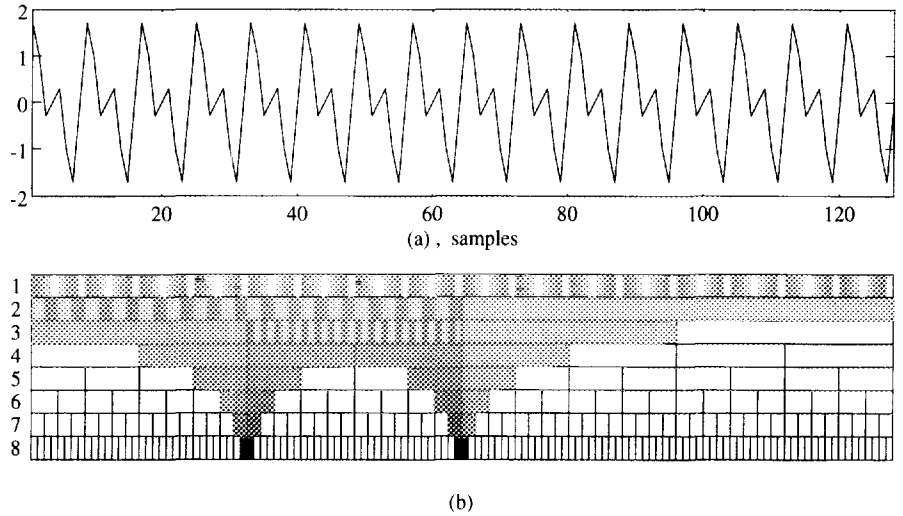

FIG. 3. The WPD of two sinusoids. (a) A frequency localized signal (b) WPD of signal. The magnitude of each element of a bin vector is displayed with black corresponding to the maximum absolute value in the tree and white corresponding to zero.

with different time–frequency characteristics will focus differently. To exploit this property for signals as in Fig. 4 we must use the full WPT and not simply the WT or a STFT in our search for features.

2.1. Energy Mapping of the Wavelet Packet Decomposition Tree

In detection terms, the formation of the WPT can be viewed as a *coherent* processing step; i.e., each sample of each signal in each bin in the full WPD can be viewed as the output of a matched filter tuned to a particular basis function. At the top of the WPD tree the basis functions are simply unit impulses at each successive time instant, and as the WPD tree is traversed downward the basis functions become more resolved in frequency and more highly decentralized in time. The matter to be determined, then, is how to use this tree of coherently processed signals to perform detection. If the signals to be detected are also described coherently, i.e., as weighted linear combinations of WPD basis functions, then a fully coherent system in which the test statistics are the same weighted linear combinations of the WPD of the received signal may be used. In the problems of interest here, however, we do *not* have such a prior model. Indeed, a fundamental premise is that the variability in these signal classes precludes such a precise representation. A second premise, however, is that the *energy* in the WPD for these signal classes does focus in a robust and useful way. This suggests a second *non-coherent*

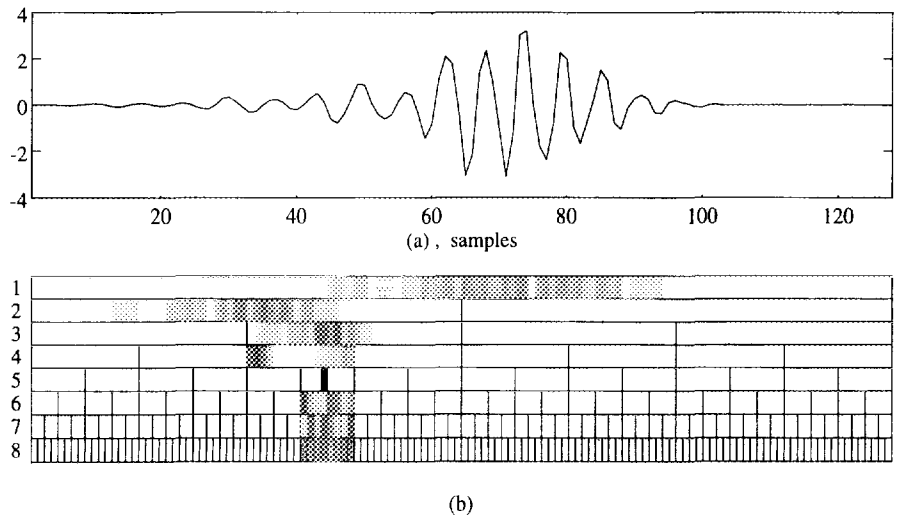

FIG. 4. The WPD of a time and frequency localized function. (a) The signal in time (b) WPD of signal.

(i.e., energy-based) processing step after the WPD has been performed. Specifically, in this work a simple energy mapping of the WPT's of the data has been done in order to begin the feature extraction process with a rudimentary exploration of signal specific characteristics.

Let $e_{\mathbf{y}}$ denote the average energy of a vector $\mathbf{y}$ having N elements. The average energy in $\mathbf{y}$ is

$$e_{\mathbf{y}} = \frac{1}{N}\mathbf{y}^T\mathbf{y}. \tag{1}$$

As an example of this energy mapping is shown in Fig. 5 for the WPD tree from Fig. 1 where a single energy value has been calculated for each bin vector. The formation of one energy value over an entire bin obviously surrenders whatever time resolution there is within each bin vector. For example, at the top of the WPD we are simply calculating total average energy, a classic test statistic in non-coherent processing.

3. THE FEATURE EXTRACTION PROCEDURE

In the formulation of a decision rule, it is desirable to find a feature set which uniquely represents each class of signals. It is generally of great importance to reduce dimensionality in order to focus information in a way that accentuates interclass distinctions and makes the task of a pattern classification scheme tractable. For example, if a neural network is used for classification, a small feature set will lessen problems that the neural network learning algorithms have with local minima. The wavelet packet feature extraction procedure was designed with the above criteria in mind.

Feature Extraction Procedure

1. Calculate the full WPT of each signal in the preclassified data set (the training set).
2. Create an energy map for each signal from its WPT.
3. Organize the energy maps into energy matrices, one for each class.
4. Calculate the singular vectors for each class.

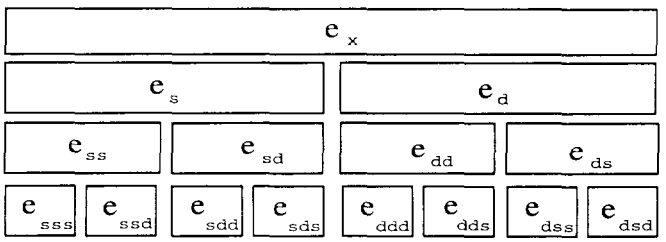

FIG. 5. Energy mapping of the WPD tree from Fig. 1.

5. Determine a parsimonious set of features from the most significant singular vectors.
6. Examine/test feature set to insure for satisfactory inter-class separation.

Steps 1 and 2 are explained in Section 2, this section details steps 3 through 6.

3.1. Matrix Representation of the Energy Maps

An energy map is found from the WPD for each of the data signals in the training set. For ease of manipulation, each energy map is represented as an energy vector by assembling the bin energies of an energy map into a column using lexicographic ordering of the bins. We number the bins from one to B and create and energy vector, $\mathbf{e}_{t,k}$, for each of our data excerpts, where the element $\mathbf{e}_{t,k}[b]$ is the energy from bin number b of the energy map for the kth signal of class t.

Next, we create a matrix for each signal class by aligning column vectors of the same class. Denote the energy matrix by E_t

$$E_t = [\mathbf{e}_{t,1} \quad \mathbf{e}_{t,2} \cdots \mathbf{e}_{t,M_t}], \tag{2}$$

where M_t represents the number of sample signals in the training set for the given class. Thus, E_t is a $B \times M_t$ matrix.

3.2. Determination of the Singular Vectors

A first step in the analysis of the transients is to quantitatively identify significant features of all energy maps from a given class. This can be done by performing the singular value decomposition (SVD) [22] of the matrices, E_t.[6]

$$E_t = U\Sigma V^T \tag{3}$$

The B-element singular vectors, $\mathbf{u}_k$, make up the columns of the $B \times B$ orthogonal matrix U. The first M_t columns of U span the column space or range of E_t.

$$U = [\mathbf{u}_1 \quad \mathbf{u}_2 \cdots \mathbf{u}_B] \tag{4}$$

The $B \times M_t$ singular value matrix, Σ, reveals the rank of E_t in the first M_t diagonal elements. The rank (or effective rank) of E_t is equal to the number of non-zero (or non-negligible) singular values.[7]

[6] The way in which to interpret singular value decomposition in the context that we (and others) have used it is as providing the Eigen-decomposition of the sample second moment matrix, $R \triangleq EE^T$.

[7] The row space and nullspace of E_t are defined in the $M_t \times M_t$ matrix, V^T. The information in V^T is not used in the analysis of the energy maps.

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{M_i} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{5}$$

These singular vectors identify the dominant energy patterns for each signal class, but what we are ultimately interested in is not only capturing the dominant energy in each class but also those features that are *distinctly different* between classes. In the next section we look for bins that have significant participation in the dominant singular vectors for one class but not for the other classes.

3.3. Construct Feature Set

We begin by finding a collection of bins that contain significant information by examination of the components of the primary singular vectors of the energy matrices. For each class, determine a singular vector to be significant if its corresponding singular value is relatively close to the maximum singular value. This will give at least one singular vector per class. Construct a preliminary feature set from the WPT bin energies that correspond to peak values of the significant singular vectors. For example, for a given singular vector, choose all bins corresponding to elements of that vector which fall above some percentage of its largest element. To form a frugal feature set, arrange features obtained from each singular vector into ancestor/descendant groupings and eliminate redundancy within a class by discarding features corresponding to the descendant bins. These steps are carefully carried out in the example of Section 4.[8]

3.4. Examine/Test Feature Set

This step is best understood through example, but a brief overview is offered here. Recall that we wish to determine the dominant bin energies that will give the best separation between classes. A preliminary validation of the feature set may be done by examination of the training set feature vector clusters and should show an inter-class variation of the features to be greater than the intra-class variation. The success of these features may be tested by using them to classify a portion of the data that has been set aside from the feature extraction process.

[8] A comment on why not to use the mean vector for extraction of features: Since $R \triangleq EE^T = P + \mathbf{m}\mathbf{m}^T$, where P is the sample covariance matrix and $\mathbf{m}$ is the mean vector of the training set, the mean vector is approximately equal to the most significant singular vector of R only in the case where $\|\mathbf{m}\|^2 \gg$ eigenvalues of P.

3.5. Common Questions

We are often asked why we don't use the simpler STFT in place of the full WPT. It should be noted that the above feature extraction procedure is performed once. After robust features are determined, the classification algorithm is simple and, if the feature set is terse, of low computational complexity. A well known trait of the WPT is that it provides a variety of tilings of the time-frequency plane; at top levels in the wavelet packet decomposition tree there is fine time resolution and coarse frequency resolution, while at subsequent levels time resolution is sacrificed for improved frequency resolution. Although any individual level of the WPT does not correspond *exactly* to a STFT with a correspondingly-sized time window, the WPT at a particular level results in the same type of tiling of the time-frequency plane as that implied by a corresponding STFT. It follows that if the optimal feature set determined by our procedure corresponded to all of the WPT bins at a single level, one would have strong evidence that the corresponding STFT would also be appropriate for generating discriminating features. However, an advantage of using the WPT is that we are not restricted to a particular fixed time–frequency tradeoff but can, in fact, identify a feature set for which each feature corresponds to a very different time-frequency tiling.

Another question sometimes raised deals with using the SVD to identify features from the energy maps. For example, why not use classification and regression trees (CART) on the energy maps to determine which features to use? This is an interesting idea, but we found our simple straightforward SVD-based procedure to yield a far more parsimonious feature extraction mechanism than would a CART-based procedure. Since no analysis of CART was done in this work, it remains an interesting possible alternative for identifying key features from the energy maps.

4. EXAMPLE

In this section, using the method described in Section 3, a feature set is found from a training set of acoustic transients. These features are used for automatic classification of a test set of these signals. The ideas presented in this paper and the following example are discussed in greater detail in Learned's thesis [13].

4.1. The Data

This example uses a collection of ocean recordings made available by the Charles Stark Draper Laboratory and the Naval Undersea Warfare Center (NUWC). The data consist of several hours of naturally occurring biologically generated underwater sounds in ambient ocean noise. The recordings have been lowpass filtered with a cutoff frequency of 5 KHz and, subsequently, sampled at 25 KHz.

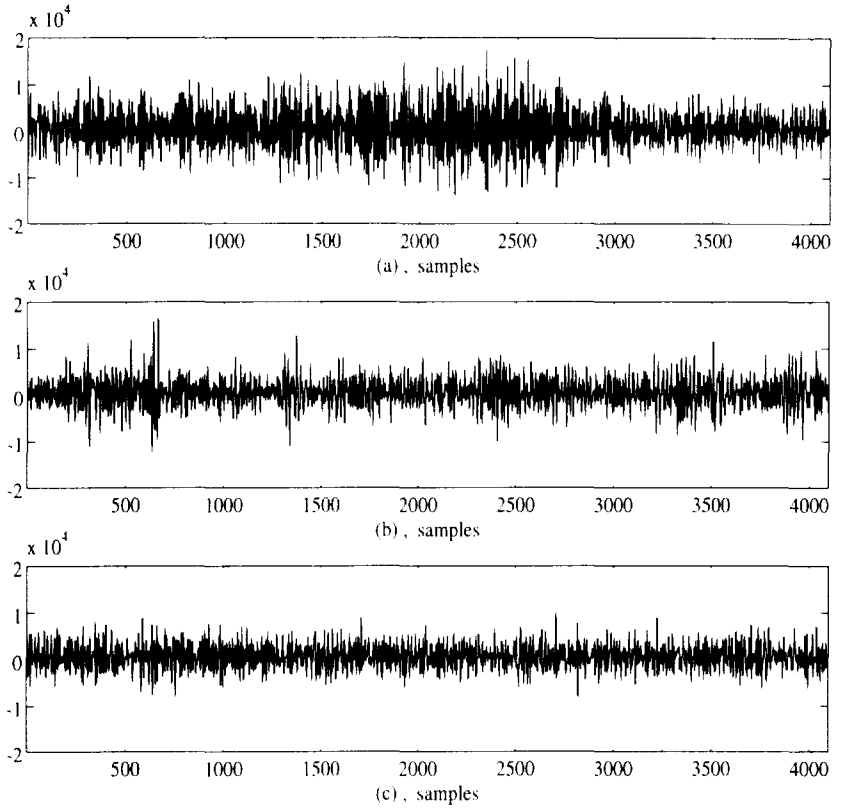

FIG. 6. Some 4096-sample (163.8 ms) excerpts from the NUWC recordings. (a) Whale click. (b) Snapping shrimp. (c) Background noise.

The biologically generated sounds are sperm whale clicks and snapping shrimp. A typical whale click has a duration of approximately 80 to 120 ms and a single snap of a shrimp has a duration on the order of 1 ms. In addition to the signals, each record contains portions of background noise alone.

A single whale click can be encompassed by a 163.8-ms or 4096-sample window which also holds one to a large number of snaps. Figure 6 shows 3 163.8-ms excerpts from the NUWC recordings, one from each class (whale click, snapping shrimp, ambient noise). We use 75 of these excerpts for the feature derivation done in Section 4.2 and 240 additional excerpts to run simulations of the classification algorithms in Section 4.3.

4.2. Find a Feature Set

The feature extraction procedure described in Section 3 was carried out on the test set of data. (1) Using our implementation of Wickerhauser's algorithm presented in [23] with the Daubechies 14 point wavelet [4], the first six levels of the wavelet packet transform of each of the 75 data excerpts were calculated. (2) An energy map was calculated for each of these 75 WPD trees. Each energy map contains $B = 63$ bin energies. (3) An energy matrix was then constructed for each class, t, where t = c (click), s (shrimp), and n (noise).

(4) Each of the energy matrices, E_c, E_s, E_n, was found to have a single dominant singular value. In particular, define the difference ratio, δ_t, between the largest and second largest singular values for each class to be

$$\delta_t = \frac{\sigma_{t,1} - \sigma_{t,2}}{\sigma_{t,1}}. \tag{6}$$

δ_t is displayed in Table 1. These values suggest that there is a single representative energy vector, corresponding to the first singular vector $\mathbf{u}_{t,1}$, for each class, t, with only a relatively small amount of variation across class members.

(5) All 63 elements (corresponding to the 63 bins of the WPD) of the primary singular vector for each class are displayed in Fig. 7. Notice that high valued elements for the noise roughly coincide with both the high valued elements for the snapping shrimp and the whale clicks. The figure

TABLE 1
Difference Ratio between the Largest and Second Largest Singular Values of the Three E_t Matrices

Class t	$\sigma_{t,1} \times 10^6$	$\sigma_{t,2} \times 10^6$	δ_t
Whale clicks	2221	285	0.87
Snapping shrimp	762	79	0.90
Background noise	412	43	0.89

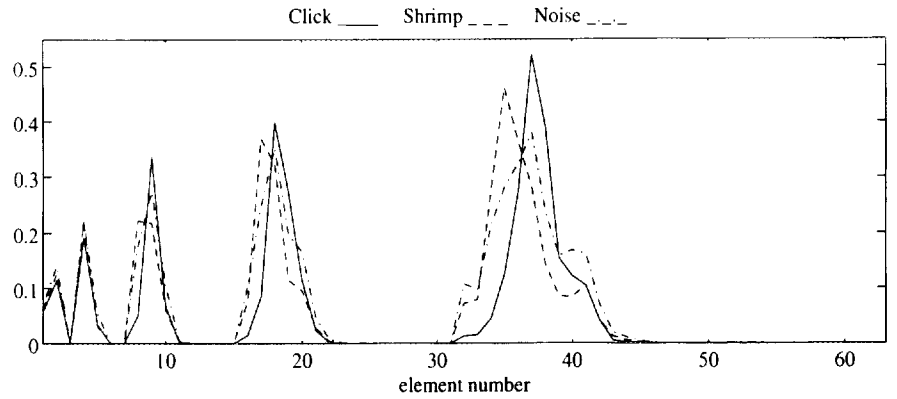

FIG. 7. Components of the 63-element primary singular vectors. Note that this plot displays elements of the three primary singular vectors with each vector representing the energy maps for its class. In other words, this plot is a lexicographical display of the representative energy bins for the three classes of energy maps.

also reveals that the high valued elements for whale clicks differ from the high valued elements for snapping shrimp. Before continuing the search for a reduced parameter feature vector from the energy maps of the training set, we may compensate for the influence of noise.

4.2.1. Compensating for the Noise. This noise normalization step does not explicitly appear in Section 3 and would fall between steps (2) and (3). Each bin energy, and, thus, each component of the energy vector, $\mathbf{e}_{t,k}$, contains both signal and noise energies. The energy maps of background noise displayed consistent energy distribution patterns. An example is seen in the singular vector of background noise in Fig. 7. This distribution of background noise energy within the energy maps may mask dominant features that may be useful in distinguishing between the shrimp and clicks. We wish to normalized each bin energy by an average noise energy so that features may be chosen without the influence of noise.

Let $\mathbf{r}$ denote the portion of the received signal vector that is due to the signal source alone. Let $\mathbf{w}$ denote the portion of the received signal vector that is due to background noise. The received signal vector, $\mathbf{x}$, may be written as a linear combination of the source signal and the background noise.

$$\mathbf{x} = \mathbf{r} + \mathbf{w}. \tag{7}$$

Let $\mathbf{x}_b$ denote the vector at bin b of the WPD of $\mathbf{x}$. Likewise, we denote the vector at bin b of the WPD of $\mathbf{r}$ and $\mathbf{w}$ as $\mathbf{r}_b$ and $\mathbf{w}_b$, respectively. Since the WPD is a linear transform, the bin vector at each bin of the WPD tree can be written as a linear combination of the bin vector due to the source and the bin vector due to the noise, i.e.

$$\mathbf{x}_b = \mathbf{r}_b + \mathbf{w}_b. \tag{8}$$

In agreement with Eq. (1), the average energy due to the bin vector $\mathbf{x}_b$ will be denoted by $e_{\mathbf{x}_b}$. Likewise, the average energy in $\mathbf{r}_b$ and $\mathbf{w}_b$ is denoted by $e_{\mathbf{r}_b}$ and $e_{\mathbf{w}_b}$, respectively. Assuming that the noise is uncorrelated with the signal allows us to write the average energy at any bin of the WPD tree as a linear combination of the average energy due to the source and the average energy due to noise.[9]

$$e_{\mathbf{x}_b} = e_{\mathbf{r}_b} + e_{\mathbf{w}_b}. \tag{9}$$

Normalization of the bin energy, $e_{\mathbf{x}_b}$, by the energy in that bin due to noise alone would give $\hat{e}_{\mathbf{x}_b}$.

$$\hat{e}_{\mathbf{x}_b} = \frac{e_{\mathbf{x}_b}}{e_{\mathbf{w}_b}} = \frac{e_{\mathbf{r}_b}}{e_{\mathbf{w}_b}} + 1. \tag{10}$$

Performing the normalization described in the above paragraphs allows for a source-signal-energy to noise-energy ratio analysis of the patterns exhibited by the energy maps.

We have already found an energy vector, $\mathbf{e}_{t,k}$, for each of our data excerpts. We now wish to find a normalized energy vector, $\hat{\mathbf{e}}_{t,k}$, for each $\mathbf{e}_{t,k}$. Element by element normalization of $\mathbf{e}_{t,k}$ by the average noise energy elements is done by

$$\hat{\mathbf{e}}_{t,k}[b] = \frac{\mathbf{e}_{t,k}[b]}{\mathbf{e}_{w,\text{ave}}[b]}, \tag{11}$$

where the element index is $b = 1, \ldots, 63$, the signal number is $k = 1, \ldots, M_t$, and each class is denoted by t = c, s, n. The average noise energy for bin b of the energy maps from our noise excerpts is used for the noise energy, $\mathbf{e}_{w,\text{ave}}[b]$. Element by element (or bin by bin) calculation of the average noise energy is done by

[9] Here, by "uncorrelated" we are in essence assuming that the time-averaged product of the noise and signal components over each bin is zero.

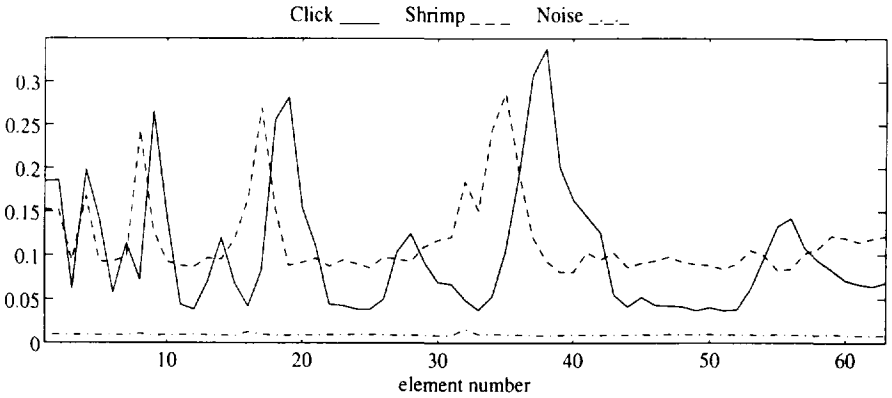

FIG. 8. Components of the 63-element primary singular vectors for the noise normalized case.

$$\mathbf{e}_{w,\text{ave}}[b] = \frac{1}{M_n}\sum_{k}^{M_n}\mathbf{e}_{n,k}[b]. \tag{12}$$

Once again, we align these normalized energy vectors into three matrices and perform singular value decomposition. The effective rank of each of these matrices was also found to be one so that one singular vector may be used as a representative energy map for each class. Figure 8 shows all 63 elements of the singular vectors found from SVD of the noise normalized energy matrices. Let us denote these noise normalized singular vectors by $\hat{\mathbf{u}}_{C,1}$ (whale clicks) and $\hat{\mathbf{u}}_{S,1}$ (snapping shrimp). We see that the high valued elements of the shrimp singular vector clearly differ from the high valued elements of the whale click singular vector and that there is no longer high valued elements for noise.[10]

4.2.2. Use Noise Normalized Energy Maps for Feature Set Determination. In forming a feature set, we look for dominant bin energies that will give us the best separation between whale clicks and snapping shrimp. We begin by finding a collection of bins that contain significant information by examination of the components of the primary singular vectors shown in Fig. 8.

We have chosen to consider a bin to be significant if the value of its corresponding element of the primary singular vector lies within 20% of the maximum component of that singular vector. The significant components of $\hat{\mathbf{u}}_{C,1}$ correspond to elements 9, 18 and 19. The significant components of $\hat{\mathbf{u}}_{S,1}$ correspond to elements 8 and 17. The two classes have no dominant bins in common. The bins corresponding to elements 8, 9, 17, 18, and 19 containing the dominant information are shaded in Fig. 9.

Reduction of the feature vector is desirable for the simplification of the decision rule, therefore, including superfluous information should be avoided. A feature set which contains a parent bin energy and all of its descendant bin energies may be redundant because any parent bin vector of the WPD tree can be constructed from its children bin vectors. Therefore, a feature set that does not incorporate both parent and child energy bins found dominant within a given singular vector is considered. Reducing the number of features used for classification will also minimize the computational complexity of the algorithm because most bins of the WPD tree will not be used and will, therefore, not be calculated.

For the whale click features, element 9 (bin $b(4, 2)$) is the parent of elements 18 (bin $b(5, 3)$) and 19 (bin $b(5, 4)$). For shrimp, element 8 (bin $b(4, 1)$) is the parent of element 17 (bin $b(5, 2)$). Noting the parent child redundancy, it

[10] We may compare the pre-noise-normalized singular vectors of Fig. 7 to the post-noise-normalized singular vectors of Fig. 8. In addition to the reduction of the noise singular vector, the normalization of the energy maps results in shifting of the locations for the peak-valued elements of the whale click singular vector.

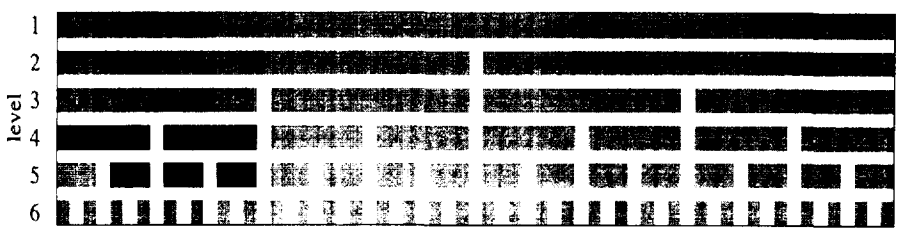

FIG. 9 The shaded bins of the energy map correspond to the dominant elements of the primary singular vectors $\hat{\mathbf{u}}_{C,1}$ and $\hat{\mathbf{u}}_{S,1}$.

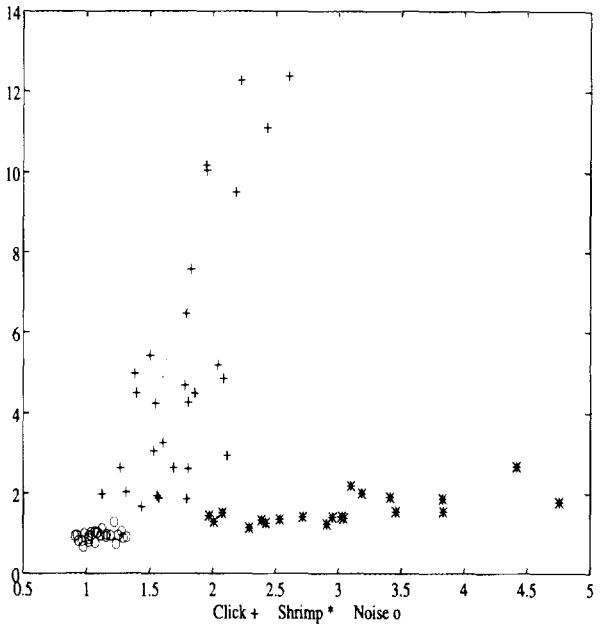

FIG. 10. Noise normalized energies from bins $b(4, 1)$ and $b(4, 2)$ of the sample energy maps that make up the training set. $b(4, 2)$ vs $b(4, 1)$.

is reasonable to see if there is enough feature separation using the energies from only the dominant parent bins at the fourth level, $b(4, 1)$ and $b(4, 2)$.

We take a look at the separation of the two signal classes provided by the two features our method identifies by showing how these features separate for the training set. Figure 10 plots the normalized energies from bins $b(4, 1)$ and $b(4, 2)$ for the energy maps of our 75 data excerpts. There is excellent separation between the click and shrimp features. Examination of the scatter plot allows us to see if our procedure has in fact determined features that clearly separate classes. That is, the features found from the training data which are to be used for classification should enhance the variation between classes and lower the variation within a class.

4.3. Tests and Results

In this section the feature extraction algorithm is tested on two groups of data. For Test I the training set is representative of the test set and we obtain excellent results using only two features. For Test II more intra-class variability is introduced by adding more shrimp data. For this appended data set the training set used for Test I is not significantly rich. In the second test we discuss what happens when an inadequate training set is used, but when we repeat the full application of the feature extraction procedure on a rich enough training set we once again find a small number features to give excellent results.

4.3.1. Test I. Once a reduced parameter feature set has been derived for a given set of sample data, a method for detection and classification must be formulated. Two pattern recognition techniques that lend themselves to the classification of signals using a training set were used in testing the utility of the wavelet-packet-based features: the nearest neighbor rule and neural networks.[11]

TABLE 2
Results Obtained from the Nearest Neighbor Rule

Number of features:	2	11
Overall classification (%)	97.92	97.92
Click classification (%)	97.87	97.87
Shrimp classification (%)	97.26	97.26
Noise classification (%)	98.63	98.63

The nearest neighbor rule, detailed by Duda and Hart in [6], uses as a training set feature vectors that have been correctly labeled with their class. A feature vector is calculated for the unknown signal. The unknown feature vector is classified with the same label as its nearest neighboring feature vector from the training set. Euclidean distance is the measure used in determining separation of feature vectors.

The neural network tests were done using the Neuralware software package [18] for building, training, and analyzing a layered neural network. A back propagation network with a tanh nonlinearity and the Widrow-Hoff-Delta Rule adaptive weighting algorithm was used in all tests.

Recall that a total of 75 signal segments (consisting of 29 whale clicks, 20 snapping shrimp excerpts, and 26 segments of noise) were used to determine the bin energies which were to be used as features. The features for this set of 75 examples were then used to establish the nearest neighbor rule and to train several neural networks. Another distinct 240 excerpts from the same overall data set were then used to test classification performance. Each test was run twice using the features determined in Section 4.2.2, once with the 2-parameter feature set comprised of the energies from bins $b(4, 1)$ and $b(4, 2)$ and once with an 11-parameter feature set comprised of the five bin energies shown in Fig. 9 plus six more adjacent bins.

The nearest neighbor rule algorithm using both the 2-parameter and 11-parameter feature sets resulted in correct classification for 97.92% of the test signals. These results are summarized in Table 2. Both nearest neighbor rule tests resulted in identical outcomes, making the same errors for both feature sets. Gaining nothing by adding more features is not surprising because the analysis done in Section

[11] Other appropriate classification algorithms may be determined by examination of scatter plots of the feature vectors. For example, linear discriminant analysis would also be effective given the type of feature clustering shown in Fig. 10.

TABLE 3
Results Obtained from the Neural Network

Number of inputs:	2	11	11
Number of ALNs in layer 1:	3	7	7
Number of ALNs in layer 2:	0	0	3
Overall classification (%)	98.33	98.75	99.17
Click classification (%)	97.87	98.44	98.44
Shrimp classification (%)	98.63	98.63	100
Noise classification (%)	98.63	98.63	98.63

4.2 determined that the energies from bins $b(4, 1)$ and $b(4, 2)$ were the dominant features necessary in distinguishing among the three classes.

Three neural networks were constructed for tests using 2 and 11 features. Excellent results were obtained for all tests. The networks (the number of adaptive linear neurons (ALN) in each layer) and their results are summarized in Table 3. The neural networks also did an excellent job; classification ranged from 98.33% to 99.17%. Here, we see that only a slight gain in performance results from the addition of the child bins to the 2-parameter feature set.

4.3.2. Test II. This test illustrates our technique on a data set for which intra-class variability is increased. This second application of our methodology emphasizes a very important point for any learning-based approach to discrimination. In particular, *any* non-parametric classification procedure needs a training set that is rich enough to encompass the full range of variability to be encountered in practice.

The range of the data set was increased by the addition of recordings of snapping shrimp taken at a different time of day and in a different region of ocean than the shrimp used in the previous sections. Testing these data with both the 2-feature and 11-feature sets (derived without samples of this new shrimp data) and the nearest neighbor rules and neural networks discussed in the previous section resulted in higher levels of incorrect classification. The reason for this can be immediately discerned from the cluster distributions shown in Fig. 11. The figure shows the two features (energies from bins $b(4, 1)$ and $b(4, 2)$) for both the new shrimp data and the original training data set. Notice that the bin energies taken from the new shrimp data records form a cluster which is distinctly separated from the bin energies for the first shrimp data set. This suggests that there is *more variability* in the bin energy patterns than that found in the first data set, requiring a richer set of features to capture this behavior. The question is, of course, whether this can be done in a way that still achieves significant feature separability *between classes.*

The feature extraction procedure was repeated on a representative training set where 16 excerpts of new shrimp data were appended to the Test I shrimp matrix and the analysis from Section 4.2 was repeated. The four largest singular values for snapping shrimp and whale clicks are shown in Table 4. We calculate a difference ratio, $\delta_t^{1,i}$, between the largest and ith largest singular values with $i = 2, 3, 4$ as shown in (13).

$$\delta_t^{1,i} = \frac{\sigma_{t,1} - \sigma_{t,i}}{\sigma_{t,1}}. \tag{13}$$

Comparing the difference ratios in Table 4 reveals $\delta_s^{1,2}$ to be relatively small, signifying that the second singular value for the shrimp class is a significant fraction of the largest singular value. Indeed, there is more variability in the energy patterns for snapping shrimp excerpts than for whale clicks. Accordingly, we expand our set of candidate features by examining two singular vectors for shrimp. Figure 12 shows the primary singular vector for whale clicks, the first singular vector for snapping shrimp, and the second singular vector for snapping shrimp scaled by $\sigma_{s,2}/\sigma_{s,1}$. This scaling was done to show the relative intensity of the two shrimp singular vectors.

We begin the search for a feature set by finding significant elements for each of the three singular vectors. We consider an element to be significant if its magnitude is within 25% of the maximum magnitude for that vector. The 13 significant values found by this procedure are marked with circles in Fig. 12 and correspond to bins at levels 4, 5, and 6 of the energy maps. From the whale click singular vector, bins $b(4, 2)$, $b(5, 3)$, $b(5, 4)$, $b(6, 6)$, and $b(6, 7)$ are significant. From the first singular vector for snapping

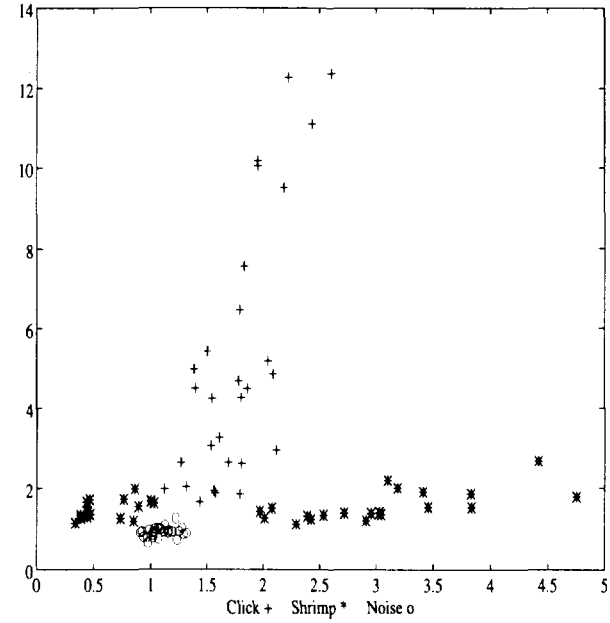

FIG. 11. Noise normalized energies from bins $b(4, 1)$ and $b(4, 2)$ of the energy maps for the original data set of snapping shrimp, whale clicks, background noise, and the new set of snapping shrimp.

TABLE 4

The Difference Ratios of Singular Values for Noise Normalized Energy Matrices That Include the New Shrimp Data

Class t	$\sigma_{t,1}$	$\sigma_{t,2}$	$\sigma_{t,3}$	$\sigma_{t,4}$	$\delta_t^{1,2}$	$\delta_t^{1,3}$	$\delta_t^{1,4}$
Snapping shrimp	65.59	25.25	13.19	9.26	0.615	0.799	0.859
Whale clicks	124.20	25.16	16.30	10.44	0.797	0.869	0.916

shrimp, bins $b(4, 1)$, $b(5, 2)$, $b(6, 3)$ and $b(6, 4)$ are significant. From the second singular vector for snapping shrimp, bins $b(5, 2)$, $b(6, 4)$, $b(6, 8)$, $b(6, 9)$, $b(6, 24)$, and $b(6, 25)$ are significant. These bins are shaded in Fig. 13a. We wish to discard all child bins from each of the three groups of features. Using only the ancestor bins within each group, we are left with seven bin energies in our feature set: $b(4, 1)$, $b(4, 2)$, $b(5, 2)$, $b(6, 8)$, $b(6, 9)$, $b(6, 24)$, and $b(6, 25)$. These seven bins are shaded in Fig. 13b.

We used the nearest neighbor rule to test the utility of these features for classifying excerpts from this more variable data set. Each rule was run twice, once with the 7-parameter feature set and once with the 13-parameter feature set. The results are summarized in Table 5. Errors made by the 13-input nearest neighbor rule are a subset of the errors made by the 7-input nearest neighbor rule. The nearest neighbor tests using both the 7 and 13 features gave excellent results ranging from 86.30% to 95.74% correct classification.

We have not presented results for neural networks for this set of experiments because of serious problems with convergence to local minima. Indeed, one of the benefits of performing the detailed feature analysis we have described is that it leads to a very small set of features that provide excellent inter-class separation. This, in turn, allows us to use a very simple classification rule, namely nearest neighbor, thus avoiding the convergence problems of neural networks.

5. CONCLUSION AND FUTURE WORK

This work has explored the feasibility of applying the wavelet packet transform to detection and classification of transient signals in background noise in the case for which the signals are not well characterized by a signal model. An adaptable systematic feature extraction procedure is presented. The features exploit signal class differences in the wavelet packet transform coefficients of pre-classified data. The formulation of a wavelet-packet-based feature set explored here combines the coherent processing of the wavelet packet decomposition with non-coherent energy calculations in each bin. From singular value decomposition of matrices made from the bin energies of wavelet packet transforms of our example data (snapping shrimp, whale clicks, and background noise) we found that only a very small number of features were necessary to distinguish among the three classes.

Specifically, four important aspects of this work are stressed. First, the approach presented is *systematic*. That is, a precise, logical procedure is describe for extracting features for signal discrimination. The procedure does not change depending upon the particular types of signals being examined (although the wavelet features the procedure identifies most certainly would). This stands in contrast to other data-based ad-hoc classification procedures which are unique to each data set for which they were intended

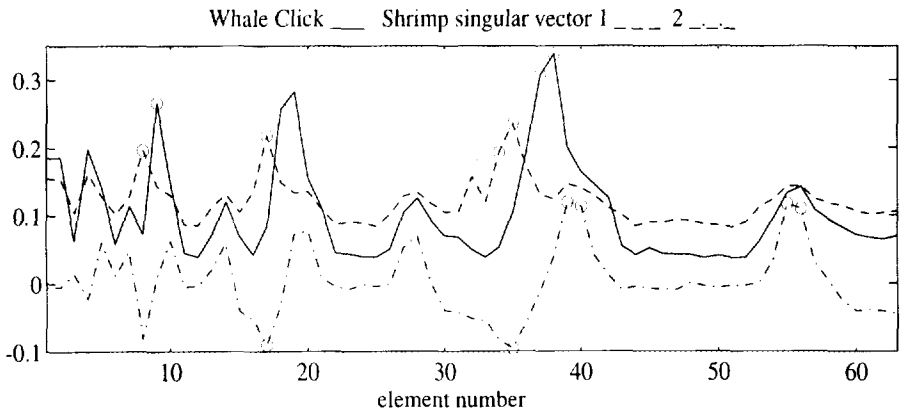

FIG. 12. The 63 elements of the primary singular vector for whale clicks and the two primary singular vectors for snapping shrimp.

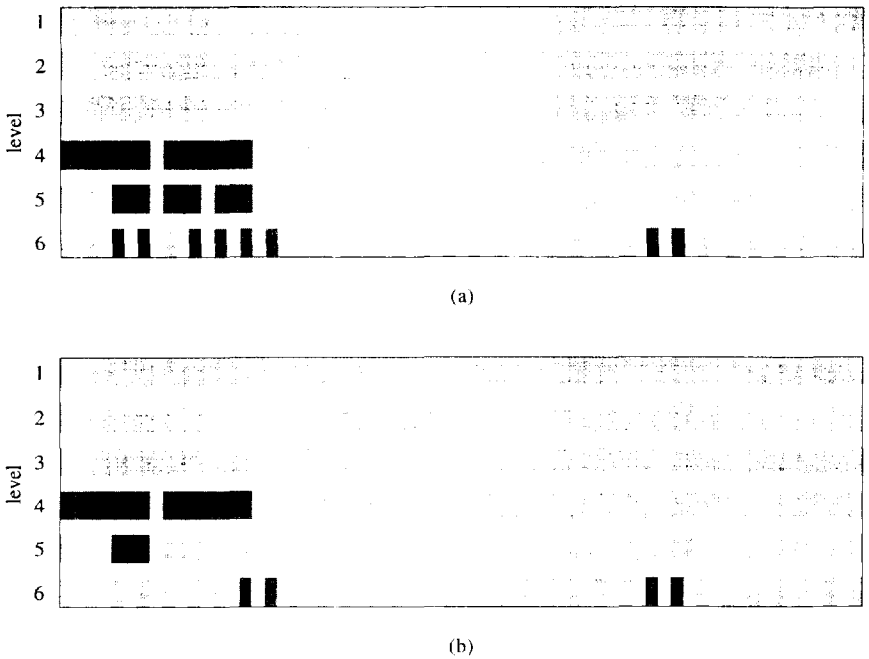

FIG. 13. The significant bins are shaded. (a) The 13 bins found significant for whale clicks and all snapping shrimp data. (b) The 7 bins of 13 that do not exhibit parent–child redundancy.

and to each person that administers them. Second, unlike other techniques using time-frequency transforms for classification, the wavelet packet transform permits automatic tuning of the features to a given signal-type. Third, and very importantly, our approach is *parsimonious* in that it attempts to identify a small number of features in which significant amounts of discrimination information is concentrated. We view this as a great strength in contrast to approaches that use vast numbers of features which are then thrown at a neural network or other black box learning algorithm. Focusing the problem down to a very small number of features allows for visualization of the discrimination problem and examination of robustness—e.g., are feature classes clearly separated? Finally, the use, promise, and success of this systematic and signal adaptable procedure has been illustrated through a particular example application. The wavelet-packet-based features obtained by our method for biologically generated underwater acoustic signals yield 86% to 100% correct classification when used as input for a neural network and a nearest neighbor rule. We believe that these results are significant not because they provide a definitive algorithm for biological acoustic transients, but rather because they provide convincing evidence that the wavelet packet transform can be used effectively as the basis for robust, systematic feature extraction and automatic identification of transient signals that cannot be well-characterized by parametric signal models.

Obviously, there is much more work that can be done to develop these ideas. First, as the results in the preceding section make clear, the development of robust classification rules require the availability of data sets that display the full range of variability present in the signal classes to be distinguished (although, as the results in Section 4.3.2 demonstrate, even a considerable level of variability may still be captured with comparatively small feature sets—a maximum of 13 in this case). Second, the choice of the wavelet used was not examined in this work and is, in general, and interesting facet to this problem. Third, a simple extension of the non-coherent energy feature calculated for each wavelet packet bin is to use a set of windowed energies for each bin, thereby enhancing temporal resolution and expanding the set of possible features considerably. The results presented here would seem to indicate that such an extension might lead to only marginal performance improvement for the application considered in this paper, but such enhanced temporal resolution may be of considerable value in other applications such as communications and active sonar/radar.

TABLE 5
Results Obtained from the Nearest Neighbor Rule in Test II

Number of features:	7	13
Overall classification (%)	91.06	95.03
Click classification (%)	94.68	95.74
Shrimp classification (%)	91.11	94.81
Noise classification (%)	86.30	94.52

ACKNOWLEDGMENTS

We express our gratitude to Professor George Verghese of the Massachusetts Institute of Technology for his contribution to this project during the early stages of the development of the feature extraction method. We sincerely thank the three referees and the editor of *ACHA* for taking the time to read and review our paper so carefully. We are grateful for their many helpful suggestions.

REFERENCES

1. T. Brotherton, T. Pollard, R. Barton, A. Krieger, and L. Marple, Application of time-frequency and time-scale analysis to underwater acoustic transients, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct. 1992.
2. T. Brotherton, T. Pollard, and D. Jones, Applications of time-frequency and time-scale representations to fault detection and classification, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct. 1992.
3. R. Coifman and M. Wickerhauser, Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory* **38,** No. 2 (March 1992).
4. I. Daubechies, Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** (Nov. 1988).
5. M. Desai and D. Shazeer, Acoustic transient analysis using wavelet decomposition, *in* "Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering, Washington, DC, Aug. 1991.
6. R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis," Wiley, New York, 1973.
7. B. Friedlander and B. Porat, Performance analysis of transient detectors based on a class of linear data transforms, *IEEE Trans. Info. Theory* **38,** No. 2 (March 1992).
8. M. Frisch and H. Messer, The use of the wavelet transform in the detection of an unknown transient signal, *IEEE Trans. Info. Theory* **38,** No. 2 (March 1992).
9. M. Frisch and H. Messer, Transient signal detection using prior information in the likelihood ratio test, *IEEE Trans. Signal Process.* **41,** No. 6 (June 1993).
10. P. Goupillaiud, A. Grossman, and J. Morlet, Cycle-octave and related transforms in seismic signal analysis, *Geoexploration* **23** (1985).
11. A. Grossman and J. Morlet, Decompositions of Hardy functions into square integrable wavelets of constant shape, *SIAM J. Math.* **15** (1984).
12. F. Lari and Z. Zachor, Automatic classification of active sonar data using time-frequency transforms, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct 1992.
13. R. Learned, "Wavelet Packet Based Transient Signal Classification," Master's thesis, Massachusetts Institute of Technology, 1992.
14. R. Loe, K. Jung, K. Anderson, S. Shen, and W. Lawton, Wavelet band features, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct 1992.
15. S. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Trans. Pattern Anal. Machine Intell.* **11** (July 1989).
16. Y. Meyer, Ondelettes et functions splines, *in* "Sem. Equations aux Derivees Partielles," Ecole Polytechnique, Paris, France, Dec. 1986.
17. F. Molinaro, F. Castanie, and A. Denjean, Knocking recognition in engine vibration signal using the wavelet transform, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct. 1992.
18. NeuralWare Inc., Penn. Center West, Pittsburgh, PA 15276, "Neural Works Professional II PLUS and Neural Works Explorer," 1990.
19. J. Nicolas, A. Lemer, and D. Legitimus, Identification automatique de bruits impulsifs en acoustique sous-marine par reseaux multicouches, *in* "International Workshop on Neural Networks and Their Applications," Nimes, France, Nov. 1989.
20. R. Priebe and G. Wilson, Applications of "matched" wavelets to identification of metallic transients, *in* "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct. 1992.
21. E. Serrano and M. Fabio, The use of the discrete wavelet transform for acoustic emission signal processing, *in* (late submission to) "Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," Victoria, BC, Canada, Oct. 1992.
22. G. Strang, "Linear Algebra and Its Applications," 4th ed., Harcourt Brace Jovanovich, 1988.
23. M. Wickerhauser, "Lectures on Wavelet Packet Algorithms," Technical report, Washington University, Department of Mathematics, 1992.