

# Optimizing Time–Frequency Kernels for Classification

Bradford W. Gillespie, *Student Member, IEEE*, and Les E. Atlas, *Member, IEEE*

**Abstract**—In many pattern recognition applications, features are traditionally extracted from standard time–frequency representations (TFRs). This assumes that the implicit smoothing of, say, a spectrogram is appropriate for the classification task. Making such assumptions may degrade classification performance. In general, any time–frequency classification technique that uses a singular quadratic TFR (e.g., the spectrogram) as a source of features will *never* surpass the performance of the same technique using a regular quadratic TFR (e.g., Rihaczek or Wigner–Ville). Any TFR that is not regular is said to be singular. Use of a singular quadratic TFR implicitly discards information without explicitly determining if it is germane to the classification task. We propose smoothing regular quadratic TFRs to retain only that information that is essential for classification. We call the resulting quadratic TFRs class-dependent TFRs. This approach makes no *a priori* assumptions about the amount and type of time–frequency smoothing required for classification. The performance of our approach is demonstrated on simulated and real data. The simulated study indicates that the performance can approach the Bayes optimal classifier. The real-world pilot studies involved helicopter fault diagnosis and radar transmitter identification.

**Index Terms**—Auto-ambiguity, classification, helicopter fault diagnosis, pattern recognition, radar transmitter identification, regular TFRs, Rihaczek, spectrogram, time–frequency kernels, tool-wear monitoring, Wigner–Ville.

## I. INTRODUCTION

QUADRATIC time–frequency representations (TFRs—referring to the quadratic class of time–frequency representations unless otherwise specified) can be uniquely characterized by an underlying function called a kernel. In past time–frequency research, kernels for a number of properties, such as finite-time support and minimizing quadratic interference, have been derived [1]. Although some of the resulting TFRs may offer advantages for classification of certain types of signals [2]–[4], the goal of sensitive detection or accurate classification is rarely an explicit goal of kernel design. Those few methods that propose to optimize the kernel for classification constrain the form of the kernel to predefined parametric functions with symmetries that may not be germane to detection or classification [5], [6].

Manuscript received January 14, 1999; revised November 1, 2000. This work was supported by a the Office of Naval Research under Grant N00014-97-1-0082. The associate editor coordinating the review of this paper and approving it for publication was Prof. Dimitrios Hatzinakos.

The authors are with the Interactive Systems Design Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195-2500 USA (e-mail: brad@ee.washington.edu).

Publisher Item Identifier S 1053-587X(01)01428-3.

Traditionally, the objective of time–frequency research is to devise a function that will describe the energy density of a signal simultaneously in time and frequency (i.e., a time–frequency *distribution*) [1]. For explicit classification, it is not necessarily desirable to represent the energy distribution of a signal in time and frequency “accurately.” In fact, such a representation may conflict with the goal of classification, generating a TFR that maximizes the separability of TFRs from different classes (here, the term class is used to refer to any grouping, arbitrary or otherwise, of “similar” data). It may be advantageous to design TFRs that specifically highlight differences between classes.

Using a standard TFR forces two (or more) transformations of the initial data to achieve robust classification. The data is transformed to a standard TFR (e.g., the spectrogram or Wigner–Ville TFR [1]), and then, a projection of the TFR to a lower dimensional space is applied (e.g., [7]). This is shown in the top panel of Fig. 1. Using a standard TFR makes implicit *a priori* assumptions about the amount and type of time–frequency smoothing required for classification. This can potentially degrade performance. Furthermore, the choice of TFR and projection algorithm must be jointly optimized; this is computationally prohibitive.

It would be better to use the optimal TFR that can be classified directly (the bottom panel in Fig. 1). We propose a method to design kernels (and thus TFRs) optimized to discriminate between predefined sets of classes. The resulting kernels are not restricted to any predefined function but, rather, are arbitrary in shape and, for certain signals, optimal. Instead of making *a priori* assumptions about the amount and type of time–frequency smoothing, our approach ascertains the necessary smoothing to achieve best classification performance. We present two optimality criteria and resulting algorithms that build on previous research in TFRs and operators [8]–[13].

To validate our approach, performance is demonstrated on simulated and real data. The simulated study shows our approach compares favorably with other techniques that have been benchmarked on this data set, approaching the performance of the Bayes optimal classifier. The two real-world pilot studies consist of helicopter fault diagnosis and radar transmitter identification. Furthermore, we experimentally validate the claim that singular TFRs will never surpass the performance of a regular TFR.

## II. BACKGROUND

A useful approach to the design of TFRs arises from a discrete Fourier transform in  $n$  (discrete time) applied to the instanta-

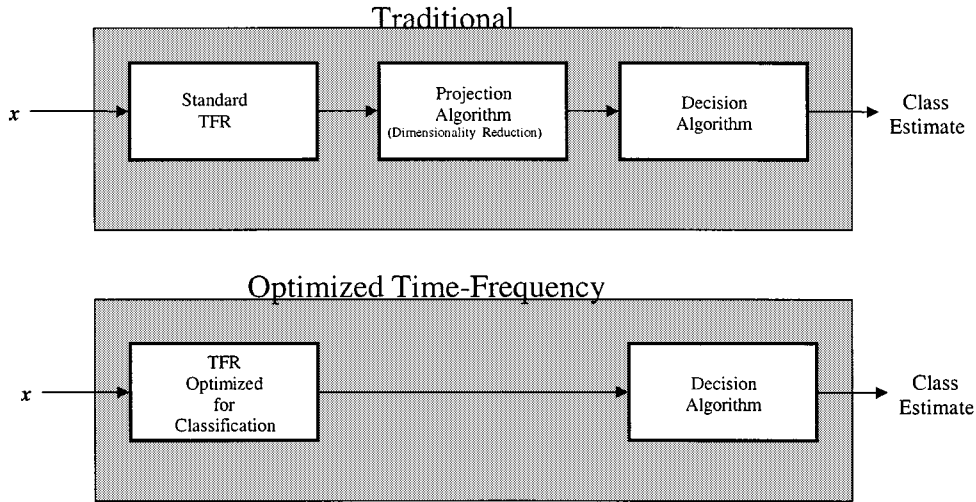


Fig. 1. Block diagram contrasting our approach to classification with traditional methods.

neous autocorrelation function  $\Re[n, \tau] = x^*[n]x[(n + \tau)_N]$ . This yields the auto-ambiguity function

$$A[\eta, \tau] = \mathcal{F}_{n \rightarrow \eta} \{ \Re[n, \tau] \} = \sum_{n=0}^{N-1} \Re[n, \tau] e^{-j(2\pi/N) n \eta} \quad (1)$$

where  $\eta$  and  $\tau$  are discrete Doppler and lag, respectively. The corresponding TFR, which is the discrete version of the Rihaczek TFR [14], is given by

$$R[n, k] = \mathcal{F}_{\eta \rightarrow n}^{-1} \{ \mathcal{F}_{\tau \rightarrow k} \{ A[\eta, \tau] \} \} \\ = \frac{1}{N} \sum_{\eta=0}^{N-1} \sum_{\tau=0}^{N-1} A[\eta, \tau] e^{-j(2\pi/N) \tau k} e^{j(2\pi/N) \eta n} \quad (2)$$

where  $k$  is discrete frequency. The characteristic function [1] of the discrete Rihaczek TFR is simply  $A[\eta, \tau]$ .

There is a kernel  $\phi[\eta, \tau]$  [15] that operates multiplicatively in both dimensions upon the auto-ambiguity function  $A[\eta, \tau]$ . The corresponding TFR is given by

$$G[n, k] = \mathcal{F}_{\eta \rightarrow n}^{-1} \{ \mathcal{F}_{\tau \rightarrow k} \{ \phi[\eta, \tau] A[\eta, \tau] \} \} \\ = \frac{1}{N} \sum_{\eta=0}^{N-1} \sum_{\tau=0}^{N-1} \phi[\eta, \tau] A[\eta, \tau] e^{-j(2\pi/N) \tau k} \\ \cdot e^{j(2\pi/N) \eta n}. \quad (3)$$

The characteristic function of  $G[n, k]$  is  $\phi[\eta, \tau] A[\eta, \tau]$ . Any nonzero extent of  $\phi[\eta, \tau]$  in  $\eta$  and/or  $\tau$  can effect a smoothing on the original Rihaczek TFR,  $R[n, k]$ , in time and/or frequency respectively.  $G[n, k]$  is a smoothed version of  $R[n, k]$ . The power of the kernel representation is that *all* TFRs can be obtained from  $R[n, k]$  by application of the appropriate kernel [1]. Thus,  $G[n, k]$  can be thought of as a generalized TFR [1].

### III. USE OF THE KERNEL AND AUTO-AMBIGUITY PLANE FOR CLASSIFICATION

The auto-ambiguity plane has very desirable properties for classification. An individual location in this plane captures

“global” information about the time frequency structure of the signal. If  $\phi[\eta, \tau] = 0$  for all values except those on the  $\eta = 0$  axis, then all temporal information is smoothed, and only steady-state frequency information is retained in the resultant smoothed TFR  $G[n, k]$ . If  $\phi[\eta, \tau] = 0$  for all values except those on the  $\tau = 0$  axis, then all spectral information is smoothed, and only temporal information is retained. Points not on either axis correspond to a sloped (nonstationary) time–frequency structure. Using the appropriate combination of weighted points, any TFR can be generated.

The ability of the aforementioned kernel to reduce time and frequency resolution, embodied within the explicit goal of classification, is the basis for the approach outlined here. When the kernel  $\phi[\eta, \tau]$  is designed with the goal of classification, we refer to it as the *signal class-dependent kernel* or, simply, *class-dependent kernel*. This kernel is denoted as  $\phi_{CD}[\eta, \tau]$ . Furthermore, we refer to the corresponding TFR  $CD[n, k]$  as the *class-dependent TFR*. This is given by

$$CD[n, k] = \mathcal{F}_{\eta \rightarrow n}^{-1} \{ \mathcal{F}_{\tau \rightarrow k} \{ \phi_{CD}[\eta, \tau] A[\eta, \tau] \} \}. \quad (4)$$

It is possible to view the class-dependent TFR and observe the time–frequency structure being exploited by the classifier.

Implicitly, the Rihaczek TFR serves as the base representation that is smoothed (since the auto-ambiguity function and the characteristic function of the Rihaczek TFR are the same). The term base representation refers to that representation to which the class-dependent kernel is applied. More generally, other TFRs could be used as the base representation. To see this generalization, define a multiplicative transformation kernel  $\phi_T[\eta, \tau]$  that transforms the Rihaczek TFR to another quadratic TFR.  $\phi_T[\eta, \tau] A[\eta, \tau]$  is the characteristic function of this new TFR. Since the transformation kernel is imposed prior to the application of the class-dependent kernel, the “new” class-dependent kernel  $\phi_{CD'}[\eta, \tau]$ , associated with the new base representation, is given by

$$\phi_{CD'}[\eta, \tau] = \frac{\phi_{CD}[\eta, \tau]}{\phi_T[\eta, \tau]}. \quad (5)$$

This kernel generates the class-dependent TFR by

$$CD[n, k] = \mathcal{F}_{\eta \rightarrow n}^{-1} \{ \mathcal{F}_{\tau \rightarrow k} \{ \phi_{CD}[\eta, \tau] (\phi_T[\eta, \tau] A[\eta, \tau]) \} \}. \quad (6)$$

Following directly from (5), the sole criterion for any candidate base TFR is that

$$|\phi_T[\eta, \tau]| \neq 0 \quad (7)$$

for all  $\eta$  and  $\tau$ , preserving all information in the original auto-ambiguity function. The TFRs that satisfy (7) are the class of regular TFRs defined by Hlawatsch [16].

One particular alternate candidate is the discrete Wigner–Ville TFR [17]. The Wigner–Ville transformation kernel is a modulating complex exponential. Hence, the Wigner–Ville TFR meets the requirements of (7). Classification performance will be equivalent to the Rihaczek TFR (the best that can be achieved using our techniques). In contrast, the spectrogram transformation kernel may be zero in regions, depending on the window being used [1]. Therefore, the spectrogram, in general, *does not* meet the requirements of (7). Thus, for arbitrary signals and classification problems, the Rihaczek and Wigner–Ville TFRs will provide superior classification (or at worst, equivalent) performance compared with the spectrogram or any other singular TFR (i.e., any TFR that is not regular).

This result is not specific to our approach. Hlawatsch notes that optimal detection cannot be performed using singular TFRs [16]. Haykin *et al.* perform dimensionality reduction on a Wigner–Ville TFR. They justify the use of the Wigner–Ville TFR by noting that it is regular; hence, the loss of information as a result of the transformation is minimized [7]. In general, any time–frequency classification technique that uses a singular TFR as a source of features will never surpass the performance of the same technique using a regular TFR. Use of a singular TFR implicitly discards information without explicitly determining if it is germane to the classification task.

Regular TFRs have not been widely accepted because they are visually difficult to interpret. Cross-terms obscure the actual time–frequency content of the signal. However, for classification, it is not necessarily desirable to represent the signal accurately in time and frequency. Good separation between classes is the sole objective. In this context, cross-terms can actually be *beneficial*. The presence or absence of a cross-term could be an excellent feature for classification.

#### IV. OUR APPROACH AND METHODS

Assume there are  $I$  training examples in the training set for a particular class. These are denoted as the  $N \times 1$  labeled vectors  $\mathbf{y}_i^{(c)}$  (i.e., the  $i$ th training example from the  $c$ th class). The goal is to design a classifier to determine the class membership of an  $N \times 1$  point vector  $\mathbf{x}$ . This is done using the training examples. The vector  $\mathbf{x}$  is not in the training set.

Theoretically, this is accomplished via a decision function  $\alpha(\mathbf{x}) \in \{1 \dots c \dots C\}$  [18]. In practice, this is usually accomplished by extracting features from a standard TFR (e.g., the spectrogram) and classifying these features directly. This makes implicit assumptions about the amount and type of

time–frequency smoothing required for classification. As we have shown, this may degrade classification performance if the standard TFR is not regular.

We propose to design and use the classifier directly in the auto-ambiguity plane. Since all TFRs can be derived from the auto-ambiguity plane, no *a priori* assumptions are made about the smoothing required for accurate classification.  $\phi_{CD}[\eta, \tau]$  is designed for each specific classification task. This design procedure uses the  $\mathbf{A}_{\mathbf{y}_i^{(c)}}$ 's [the matrix representation of the auto-ambiguity function derived from  $\mathbf{y}_i^{(c)}$ ] to directly determine the required smoothing.  $\mathbf{x}$  is classified via  $\alpha(\phi_{CD} \circ \mathbf{A}_{\mathbf{x}}) \in \{1 \dots c \dots C\}$ , where  $\circ$  is an element-by-element product.

Our approach to kernel design and classification is a generalization of the signal class-dependent method described in more detail before [9]–[11]. A brief overview of this previously described approach is provided as it is essential to understanding the motivations for our modifications. We extend the class-dependent methodology to a general kernel for classification.

##### A. Previous Approach

The previously described approach [9]–[11] finds the single kernel  $\phi_{CD}$  that maximizes the distance, in a mean-square sense, between the average smoothed TFRs for each of  $C$  different classes. This average smoothed TFR is defined as

$$\bar{\mathbf{G}}^{(c)} = \frac{1}{I} \sum_{i=1}^I \mathbf{G}_{\mathbf{y}_i^{(c)}}. \quad (8)$$

$\mathbf{G}_{\mathbf{y}_i^{(c)}}$  is the matrix representation of the generalized TFR derived from  $\mathbf{y}_i^{(c)}$ . The class-dependent kernel is given by

$$\phi_{CD} = \arg \max_{\phi} \left\{ \sum_{i=1}^C \sum_{j=1}^C \left\| \bar{\mathbf{G}}^{(i)} - \bar{\mathbf{G}}^{(j)} \right\|_F^2 \right\} \quad (9)$$

where  $\|\cdot\|_F^2$  is the Frobenius norm of the matrix. McLaughlin and Atlas [11] have shown that the kernel that achieves this maximization can be obtained directly in the auto-ambiguity plane. This is given by

$$\phi_{CD} = \arg \max_{\phi} \left\{ \sum_{i=1}^C \sum_{j=1}^C \left\| \phi \circ \bar{\mathbf{A}}^{(i)} - \phi \circ \bar{\mathbf{A}}^{(j)} \right\|_F^2 \right\}. \quad (10)$$

$\bar{\mathbf{A}}^{(c)}$  is the “average” auto-ambiguity function of class  $c$ , which is defined as

$$\bar{\mathbf{A}}^{(c)} = \frac{1}{I} \sum_{i=1}^I \mathbf{A}_{\mathbf{y}_i^{(c)}}. \quad (11)$$

A unitary energy constraint is imposed on the kernel given by

$$\|\phi_{CD}\|_F^2 = 1. \quad (12)$$

The kernel that results from the maximization in (10), constrained by (12), is “1” at the single location in  $\eta, \tau$ , where the  $\bar{\mathbf{A}}^{(c)}[\eta, \tau]$ s are most separated and “0” elsewhere. McLaughlin

and Atlas [11] have shown that this corresponds to the kernel that maximizes (9), which is the class-dependent kernel.

In practice, class-dependent kernel design is accomplished by rank ordering the kernel points according to the mean separation between classes and selecting the highest ranked point. This kernel is set to “1” at a single location corresponding to the highest ranked point. For actual classification of an unknown series  $\mathbf{x}$ , the auto-ambiguity function corresponding to the unknown example  $\mathbf{A}_x$  is multiplied, in  $\eta$  and  $\tau$ , by the previously determined binary kernel mask.  $\mathbf{x}$  is assigned to the nearest class in Euclidean distance.

A particular class-dependent kernel is optimized to separate the average TFRs of  $C$  specific classes. This kernel, in general, will not extend to other sets of classes. A new kernel must be estimated for each particular classification problem.

### B. Toward a Bayes Optimal Kernel

The class-dependent kernel optimized by the mean-square distance criterion is inadequate to handle the wide range of within-class variance seen in real-world applications. The above outlined approach “selects” only that point (in  $\eta$  and  $\tau$ ) that maximizes

$$\sum_{i=1}^C \sum_{j=1}^C \left| \bar{A}^{(i)}[\eta, \tau] - \bar{A}^{(j)}[\eta, \tau] \right| \quad (13)$$

which is the location in the auto-ambiguity plane with maximally separated class means. This corresponds to maximally separate *average* TFRs. For accurate classification, not only must classes be well separated in the mean, but the within-class variance must be small relative to this separation. Furthermore, “selecting” multiple kernel points is essential for adequate classifier performance over a broad range of signals. For example, the linear combination of two dimensions may provide perfect class separation that cannot be achieved in either dimension alone.

Here, we propose two kernels for classification. The first uses a linear discriminant function that provides superior discriminatory power, compared with the original class-dependent approach. Often, the solution to this kernel is ill posed. We propose a second approach to the design of class-dependent kernels. This alternate approach allows classification of limited size data sets, with minimal performance loss compared with the linear discriminant kernel. It is important to note that using either of these approaches no longer guarantees that the resulting TFRs satisfy (9). However, the resulting TFRs are superior in terms of classification performance, and for certain signals, these kernels can be optimal for classification.

*Approach 1—Linear Discriminant Kernel:* In the linear discriminant kernel, points can take all possible complex values, provided the unitary energy constraint (12) is satisfied. The kernel is optimized using estimated classification error or related measures. A computationally simple and mathematically tractable approach is to assume a linear discriminant classifier and optimize the kernel by the mean square error criterion [18]. The resulting kernel is optimal for classification, provided the underlying points in the kernel are Gaussian and the covariance matrices of each class are equal.

With a linear discriminant function, the classification rule estimates the class of the unknown example based on the value of  $\|\phi_{CD} \circ \mathbf{A}_x\|_F^2$ . This expression is the matrix extension of standard linear discriminant analysis [18]. Points in the auto-ambiguity function take the role of features, and the kernel takes the role of the weights. This has an interesting interpretation in the time–frequency domain. The optimal kernel separates classes by energy in the auto-ambiguity plane (or equivalently the time–frequency plane).

Previously, the criterion for the class-dependent kernel was defined explicitly [see (10)] under the unitary energy constraint. This criterion proved inadequate for actual classification. This raises the following question: What criterion is being satisfied with the linear discriminant kernel presented here? The kernel is trying to minimize the expected classification risk. The minimum classification risk is called the Bayes risk [18]. In general, it is difficult to optimize the kernel using the minimum Bayes risk criterion; our intention is to design a kernel that approximates the minimum Bayes risk kernel. In fact, the linear discriminant kernel approaches a minimum mean-squared-error approximation to the Bayes discriminant function [18].

*Approach 2—Fisher’s Discriminant Ratio Kernel:* Determining the linear discriminant kernel involves estimating the  $N^2$  weight parameters corresponding to each point in the auto-ambiguity plane. In many situations, this problem is ill posed.  $N^2$  is often greater than the number of examples at our disposal. Buckheit and Donoho refer to this condition as the “neo-classical setting” of linear discriminant analysis [19]. This occurs most often when applying this technique to real-world data, where the number of examples is limited. This makes the estimated kernel values (i.e., linear discriminant weights) unreliable at best [20]. We propose an approach to cope with limited training examples. As a benefit, the classifier is not restricted to a linear discriminant function. This flexibility is incorporated into the overall structure by separating the kernel design and classification steps. To accomplish this, values of the kernel are constrained such that

$$\phi_{CD}[\eta, \tau] \in \{0, K^{-1/2}\} \quad (14)$$

where  $K$  is the number of nonzero points in the class-dependent kernel. This ensures the unitary energy constraint is satisfied. The class-dependent kernel is selecting, in effect, “features” from the set of points that make up the auto-ambiguity function. These “features” can be used to estimate the class of the unknown example using any classifier architecture. These two distinct steps (kernel design and classification) will be discussed here in more detail.

*Kernel Design (Feature Ranking):* To closely approximate the linear discriminant kernel, we find the points in the kernel that maximize the Fisher’s discriminant ratio (FDR) given by

$$\text{FDR}[\eta, \tau] = \frac{\sum_{i=1}^C \sum_{j=1}^C \left| \bar{A}^{(i)}[\eta, \tau] - \bar{A}^{(j)}[\eta, \tau] \right|^2}{\sum_{i=1}^C (\sigma^{(c)}[\eta, \tau])^2} \quad (15)$$

where

$$\sigma^{(c)}[\eta, \tau] = \frac{1}{I} \sum_{i=1}^I \left| A_{\mathbf{y}_i^{(c)}}[\eta, \tau] \right|^2 - \left| \bar{A}^{(c)}[\eta, \tau] \right|^2. \quad (16)$$

$\sigma^{(c)}[\eta, \tau]$  is the estimated standard deviation at a particular point in  $\eta$  and  $\tau$  for the  $I$  training examples from class  $c$ . FDR provides a rank ordering of kernel points for classification. The main assumption of FDR is that the underlying probability distribution of each kernel location (feature) is Gaussian or at least unimodal [21]. If this assumption is satisfied, FDR is maximized when the separation between means of the class clusters is large and the within-class variance is small. Notice that if every location in the auto-ambiguity plane is Gaussian distributed with equal variance and  $K = 1$ , then this technique defaults to the original class-dependent approach.

The optimal number of nonzero points is determined by evaluating the classifier performance using the  $K$  best kernel points (i.e., the  $K$  points with the largest Fisher's discriminant ratio).  $K_{opt}$  is selected to be the  $K$  for which the probability of correct classification is greatest.

For many signals, a significant amount of correlation exists in the auto-ambiguity plane. FDR ranks features in a single dimension, failing to account for this correlation. Correlation often improves overall classification performance and, thus, should be accounted for. More sophisticated techniques rank-order points in multiple feature dimensions (e.g., Fisher's linear discriminant function [21] and Procrustes angle [22]). Ranking in this manner accounts for the potential gains afforded by exploiting the correlation in the selected points (features). Unfortunately, the number of points in the kernel is often greater than the number of examples, making standard multidimensional feature ranking unreliable. As a compromise, we exclude those kernel points that are strongly correlated with higher ranked kernel points (using FDR). These highly correlated points do not contribute to classification [21]. We have found that this suboptimal approach improves the overall classification performance compared with FDR alone.

*Classification:* After designing a kernel using  $I$  examples from each of the  $C$  classes, actual classification is performed. Given a particular unknown  $N \times 1$  test signal vector  $\mathbf{x}$  (the classifier is not trained on this example), the classifier estimates the class membership of this example. Assume each segment belongs to only one of  $C$  possible classes. To classify  $\mathbf{x}$ , an  $N \times N$  point auto-ambiguity function  $\mathbf{A}_{\mathbf{x}}$ , centered about the origin, is estimated from the signal.

The class of  $\mathbf{x}$  can be estimated by standard classification techniques (e.g., linear or quadratic discriminant functions, neural networks, or classification trees). We elect to classify via a quadratic discriminant function given by

$$\alpha(\phi_{CD} \circ \mathbf{A}_{\mathbf{x}}) = \arg \min_c \{l_c(\phi_{CD} \circ \mathbf{A}_{\mathbf{x}})\} \quad (17)$$

where  $l_c(\phi_{CD} \circ \mathbf{A}_{\mathbf{x}})$  is the likelihood function for class  $c$  and is given by

$$\begin{aligned} l_c(\phi_{CD} \circ \mathbf{A}_{\mathbf{x}}) = & \left( \phi_{CD} \circ \mathbf{A}_{\mathbf{x}} - \phi_{CD} \circ \bar{\mathbf{A}}^{(c)} \right)^T \\ & \cdot \Sigma_c^{-1} \left( \phi_{CD} \circ \mathbf{A}_{\mathbf{x}} - \phi_{CD} \circ \bar{\mathbf{A}}^{(c)} \right) \\ & + \ln |\Sigma_c| - 2 \ln \pi_c \end{aligned} \quad (18)$$

where  $(\cdot)^T$  denotes matrix transpose and  $|\cdot|$  the determinant. The *a priori* probability of class  $c$ ,  $\pi_c$  is assumed known. The class mean  $\bar{\mathbf{A}}^{(c)}$  and class covariance matrix  $\Sigma_c$  are estimated from the training data. Typically, for our applications, the *a priori* probability is equal for all classes, and thus, the last term in (18) can be neglected. In practice, the dimensions that are set to zero by the class-dependent kernel are removed before computing the likelihood function.

### C. Optimization of Multiple TFRs

In the framework of class-dependent kernels, it is possible to jointly optimize multiple kernels. Each kernel is derived from a different signal, which is referred to as a channel. This has utility in situations where multiple sensors capture information about the same event. In the linear discriminant framework, the kernels are determined by computing the transformation weights on all the distributions jointly. The length of the input signal from each channel and/or the size of the associated auto-ambiguity function need not be the same. As an added benefit, this approach allows the importance of individual channels to be examined. Unimportant channels will have an all-zero class-dependent kernel.

This procedure is readily extended to FDR. Points in the auto-ambiguity functions are jointly ranked using FDR. Classification proceeds as described previously.

### D. Visual Appearance

Because accurate classification, and not visual appearance, is our sole design criterion, the class-dependent TFR often does not provide a visually satisfying appearance. The kernel smoothes  $R[n, k]$  "globally." In general,  $K$  is small (due to the "curse of dimensionality" [18]), forcing a severe smoothing of  $R[n, k]$ . For example, if two classes are distinguished by the presence of a transient at time  $n_0$ , the class-dependent TFR may not correspond to a broadband impulse at  $n_0$ . Rather, the average TFR for each class will be a globally smoothed version of the original TFR.

## V. EXPERIMENTAL RESULTS

The performance of the class-dependent approach is explored on a variety of data sets. These experiments will demonstrate the ability of our proposed approach to *automatically* determine the amount and type of time-frequency information required for accurate classification. A simulated data set is used to compare our approach to the Bayes optimal classifier. This data set has the added benefit of allowing the performance of the linear discriminant and Fisher's discriminant ratio kernels to be explored. Two pilot data sets are studied to benchmark our approach under real-world conditions. Droppo and Atlas give additional real-world results, using this technique, on the task of vowel discrimination [13].

### A. Simulated Results

The use of a simulated data set allows analysis of the performance of the original class-dependent formulation, the Fisher's discriminant ratio kernel, and the linear discriminant kernel. Most real-world data sets do not contain enough examples for

the linear discriminant kernel to be utilized. With this simulated data set, all three techniques can be compared with the Bayes optimal classifier. The data set used was originally proposed by Breiman *et al.* [23] and is considered a difficult pattern recognition problem [19]. The waveforms that define each class are

$$\text{Class 1 } x[n] = \mathcal{U}h_1[n] + (1 - \mathcal{U})h_2[n] + \mathcal{N}[n]$$

$$\text{Class 2 } x[n] = \mathcal{U}h_1[n] + (1 - \mathcal{U})h_3[n] + \mathcal{N}[n]$$

and

$$\text{Class 3 } x[n] = \mathcal{U}h_2[n] + (1 - \mathcal{U})h_3[n] + \mathcal{N}[n]. \quad (19)$$

$\mathcal{U}$  is a number uniformly distributed over the interval (0, 1),  $\mathcal{N}[n]$  is a zero mean normal random variable with variance 1, and  $n = 1 \cdots 21$ . The functions  $h_i[n]$  are shifted triangle waveforms defined as

$$h_1[n] = \max(6 - |n - 7|, 0)$$

$$h_2[n] = h_1[n - 8]$$

and

$$h_3[n] = h_1[n - 4]. \quad (20)$$

Breiman *et al.* have determined that the Bayes error rate for this example is 0.14 (i.e., the optimal classifier will correctly estimate the class of a random unknown example  $\mathbf{x}$  86% of the time).

Each classifier was trained on 3000 examples (1000 per class). Three thousand independent examples (1000 per class) comprised the test set. This process was repeated ten times, and the results were averaged to provide an estimate of classifier performance. To design the Fisher's discriminant ratio kernel, a correlation threshold of 0.95 was used.

The performance of the three techniques is shown in Table I along with a number of other techniques that have been benchmarked on this data. The linear discriminant kernel outperformed other techniques with an average error rate of 0.149 (0.009 worse than the Bayes optimal classifier). Our proposed Fisher's discriminant ratio kernel performed nearly as well as the linear discriminant kernel, with an overall misclassification rate of 0.155. This indicates that the performance loss incurred by the Fisher's discriminant ratio kernel, while measurable, is acceptably small. The original class-dependent approach performed significantly worse, with a misclassification rate of 0.254.

These results show that the generalizations of the class-dependent approach presented here provide superior discriminatory power compared with the original formulation, approaching the performance of the Bayes optimal classifier. Furthermore, our approach compares favorably with other techniques that have been benchmarked on this data set.

### B. Helicopter Gearbox Monitoring

An interesting pattern recognition application is the monitoring of mechanical systems, specifically helicopter gearboxes, for wear and/or imminent failure. Early detection of failure can prevent costly repairs and help avoid catastrophic failure of the helicopter in mid-flight. Currently, gearboxes are dismantled and rebuilt at scheduled intervals to avoid problems. To simplify this process, automated helicopter gearbox monitoring

TABLE I  
ERROR RATE OF THE VARIATIONS OF THE CLASS-DEPENDENT APPROACH ON THE SIMULATED DATA SET. THE BAYES ERROR RATE IS GIVEN FOR COMPARISON. RESULTS FOR THE FINAL SIX TECHNIQUES WERE TAKEN FROM THE ASSOCIATED REFERENCES

Method	Error Rate
Bayes optimal classifier [23]	0.140
Linear discriminant kernel — described herein	0.149
Fisher's discriminant ratio kernel — described herein	0.155
Original class-dependent methodology [9], [10], [11]	0.254
Wavelet packets [19]	0.150
Mixture discriminant analysis [25]	0.157
Cosine packets [19]	0.168
Linear discriminant analysis [25]	0.191
Flexible discriminant analysis/multivariate adaptive regression splines (degree=1) [24]	0.191
Tree structured classifier (with single coordinate splits) [23]	0.289

systems have been developed. Accelerometers are mounted at various locations on the gearbox. Vibration patterns from these accelerometers are, in general, analyzed by using time-domain (e.g., [26]), frequency-domain (e.g., [27]), or combined time-frequency (e.g., [2], [28]) approaches. In the latter case, the application of time-frequency or related techniques have used standard TFRs that may degrade classification performance.

Instead of imposing *a priori* smoothing in time or frequency, we utilize the class-dependent methodology to determine the best smoothing directly from the data. For example, if transients occurring at a certain phase of rotation were important for detecting imminent failure due to crack formation, it would be best to have high resolution in time. Alternatively, if a specific harmonic were important for detecting root fatigue, it would be best to have high resolution in frequency.

Data provided by the Applied Research Laboratory (ARL) at Pennsylvania State University was used for classification [29]. These data were collected utilizing Westland Helicopters Ltd.'s Universal Transmission Test Rig to study a CH46 aft transmission. Faults were individually induced into the gearbox. Eight accelerometer time series were collected for each condition at varying torque levels. Four faults and an associated normal (no fault) condition were collected at four separate torque levels (27%, 50%, 70%, and 100% torque). Class labels were assigned as follows:

- Class 1—Input pinion bearing corrosion.
- Class 2—Spiral bevel input pinion spalling.
- Class 3—Collector gear crack propagation.
- Class 4—Quill shaft crack propagation.
- Class 5—No fault.

The multidimensional time series for each class was divided into 2048-point, nonoverlapping, and contiguous segments, with eight channels per segment. Prior to classification, each segment is demeaned (i.e., mean removed) and is normalized to a standard deviation of one. The total number of segments for each fault was 800 (200 at each torque level), where each was 19.87 ms long. The system is trained using all available

TABLE II

ERROR RATE OF THE CLASS-DEPENDENT APPROACH ON THE TASK OF HELICOPTER GEARBOX MONITORING. THE CLASSIFIER WAS TRAINED ON DATA COLLECTED AT 27% AND 100% TORQUE. TESTING WAS PERFORMED ON DATA COLLECTED AT 50% AND 70% TORQUE. (a) CONFUSION MATRIX FOR THE 50% TORQUE EXAMPLE. (b) CONFUSION MATRIX FOR THE 70% TORQUE EXAMPLE. THE FISHER'S DISCRIMINANT RATIO KERNEL WITH THE RIHACZEK BASE REPRESENTATION IS EMPLOYED. FIVE OF THE EIGHT ACCELEROMETERS WERE USED FOR CLASSIFICATION. SEE THE TEXT FOR DETAILS

		Estimated Class				
		Class 1	Class 2	Class 3	Class 4	Class 5
True Class	Class 1	1.00	0	0	0	0
	Class 2	0	0.96	0	0	0.04
	Class 3	0	0	1.00	0	0
	Class 4	0	0	0	1.00	0
	Class 5	.005	0	0	0	0.995

(a)

		Estimated Class				
		Class 1	Class 2	Class 3	Class 4	Class 5
True Class	Class 1	0.915	0	0	0.085	0
	Class 2	0	1.00	0	0	0
	Class 3	0	0	1.00	0	0
	Class 4	0	0	0	1.00	0
	Class 5	.115	0	0	0	0.885

(b)

data collected at the 27% and 100% torque levels. Testing is then performed on data collected at the 50% and 70% torque levels. The goal is to determine, from a 19.87-ms segment of data at torque levels unseen by the classifier, the class of the segment. This experiment demonstrates several points of our proposed approach:

- 1) the ability to automatically determine the amount and type of time-frequency information that is important for classification;
- 2) the ability to determine automatically important channels (in this case accelerometers) for classification;
- 3) the ability to generalize to torque levels that have not been observed by the classifier.

Using the proposed Fisher's discriminant kernel with the Rihaczek base representation, an average correct classification rate of 0.976 was achieved on the combined 50% and 70% test sets. Table II shows the confusion matrices for both the 50% and 70% torque examples. It is important to realize that the results reported here reflect decisions made at 19.87-ms intervals. If longer averaging were used, the error rate could be dramatically reduced.

Fig. 2 shows a plot of the average error rate as a function of kernel points. Best performance is achieved with 13 kernel points; if the dimensionality of the kernel is increased beyond 13, performance degrades. These 13 kernel points are distributed across five accelerometer channels. Consequently, our approach has automatically determined that three of the eight accelerometers did not contain information useful for accurate classification.

Our approach has automatically determined that only stationary information is required for accurate classification of this

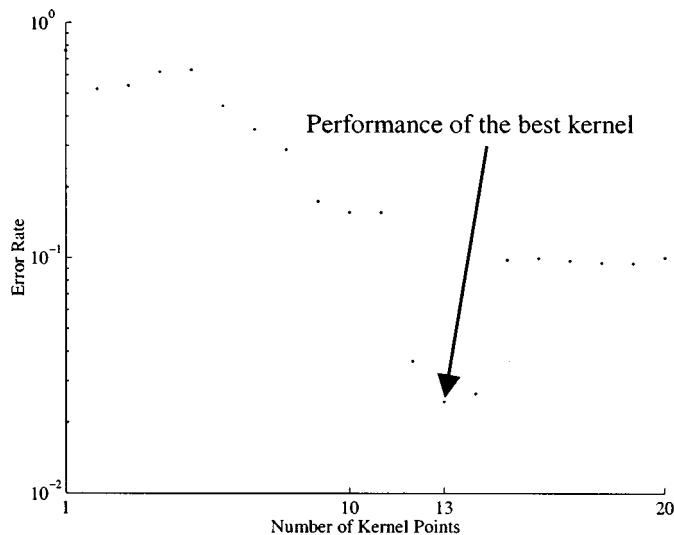


Fig. 2. Rank-order curve of the average error on the helicopter gearbox experiment as a function of points in the kernel. The minimum error rate is achieved with 13 kernel points. Increasing the number of kernel points beyond this point (i.e., decreasing the smoothing in time and frequency) decreases the overall performance of the system.

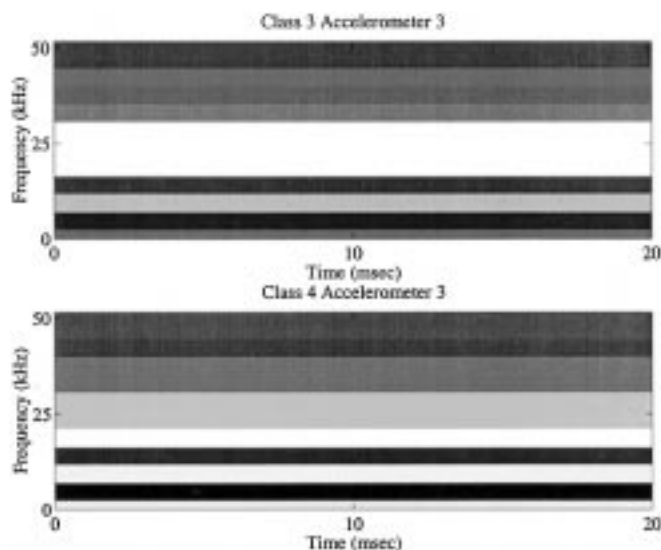


Fig. 3. Two class-dependent time-frequency representations from accelerometer 4 of the helicopter gearbox data. These represent the smoothed TFRs from two different faults. These TRFs represent only one of the five accelerometers used for classification.

data set. Only points on the  $\eta = 0$  axis (i.e., the autocorrelation sequence) were selected for classification. Fig. 3 shows two examples of the class-dependent TFR for the helicopter data. Clearly, only stationary frequency information has been retained. These examples compare two classes on a single accelerometer. For actual classification, five accelerometers were utilized. Consequently, more information is being used for classification that is being shown in Fig. 3.

Since only stationary spectral information was needed for accurate classification, standard spectral analysis techniques would be expected to perform at least as well as our technique and would be simpler and more efficient for on-line implementation. The value of our technique in this situation is 1) to conclusively show that nonstationary information was not

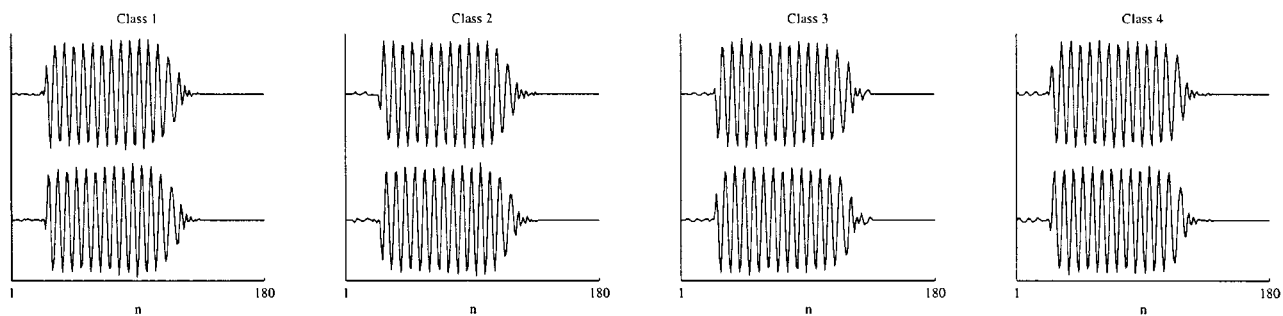


Fig. 4. Example signal from each class. (Top) In-phase and (bottom) quadrature.

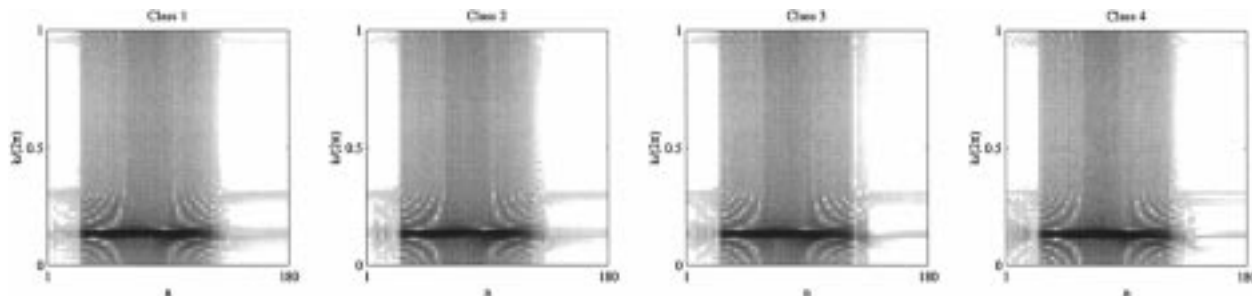


Fig. 5. Log magnitude of the original unsmoothed TFRs that correspond to the signals shown in Fig. 4. The largest magnitude is represented by the darkest gray-scale value.

germane to this classification problem and 2) to determine, automatically, an optimal combination of accelerometers. In the following section, we present a problem where nonstationary information is essential. Thus, our technique would be suitable for on-line implementation on this problem.

### C. Radar Transmitter Identification

One application of interest in radar signal processing is the detection and classification of individual radar signals. The goal of radar transmitter identification is to determine the particular transmitter from which a signal originated, using *only* the received waveform. No localization information is exploited to accomplish this task. Each transmitter must be identified in the presence of other transmitters of the same type (i.e., same model number but different serial number). Individual transmitter identification can be accomplished by exploiting the unintentional modulation present in these radar signals. This modulation is a result of subtle variations between particular transmitter components and acts as a signature for an individual radar station.

A variety of techniques could be used to identify individual transmitters through the unintentional modulation present on the radar signal. Instead of imposing *a priori* a time, frequency, or combined time–frequency approach, it is better to make no assumptions and determine the required smoothing directly from the data. If the center frequency were important for transmitter identification, it would be best to have high resolution in frequency and little or no resolution in time.

Data provided by the Naval Research Laboratory (NRL) is utilized in this work [30]. This data set contains ten radar pulses from four transmitters. This data comprises three tests from each of the four sources called A2, CCC2, F2, and H2. These will be

denoted as class one through four, respectively. Each pulse contains 180 complex samples (i.e., in-phase and quadrature components). An example of a radar pulse from each class is shown in Fig. 4 and the corresponding log magnitude Rihaczek TFR in Fig. 5.

Prior to classification each segment is demeaned and normalized to a standard deviation of one. This prevents classification based on irrelevant or variable features. The second step is an outgrowth of the class-dependent approach. The center frequency of the transmitter is a variable parameter. Because our method seeks to find a time–frequency representation that maximizes between-class separation, if a particular class in the training set contains a center frequency bias, this will be used as an essential class discriminator. The variability of the center frequency makes this unusable as a discriminatory feature. There are three possible solutions to ensure that this feature is not incorporated into the classifier.

- Given enough representative data from each transmitter (presumably including variability in the center frequency), the classifier will discard this feature as a possible means of classification. This is equivalent to the classifier “learning” that the center frequency is irrelevant.
- Only the magnitude of the radar pulse is used for classification. This presumes that there is enough information in the envelope of the radar signature to discriminate classes.
- The data set is preprocessed to modulate all pulses to the same center frequency. This involves estimation of the center frequency of each pulse and modulation to a new predetermined center frequency.

Due to the size of the data set provided, the latter method is preferred. The large SNR of this data makes estimation of the center frequency of the signal relatively easy. It was determined that 34 out of the 40 examples had a center frequency of

TABLE III

ERROR RATE OF THE CLASS-DEPENDENT APPROACH USING A NUMBER OF BASE REPRESENTATIONS ON RADAR TRANSMITTER IDENTIFICATION. ONE THOUSAND EXAMPLES PER CLASS WERE USED TO ESTIMATE CLASSIFIER PERFORMANCE

Case	Class-Dependent Base Representation			
	Rihaczek	Wigner-Ville	Narrowband Spectrogram	Wideband Spectrogram
1	0	0	0	0
2	0.0008	0.0008	0.0008	0.0043
3	0.0495	0.0495	0.0495	0.0900
4	0.0760	0.0760	0.0760	0.1190

0.151( $2\pi$ ) radians per sample, whereas six pulses (all from class one) had a center frequency of 0.145( $2\pi$ ) radians per sample.

The selected preprocessing algorithm for this data was to modulate all signals to a center frequency of 0.151( $2\pi$ ) radians per sample. Once this preprocessing algorithm is applied to the data, transmitter identification is implemented as described above.

In order to experimentally study the class-dependent approach,  $N$ -fold cross-validation is used [23]. The data are randomly divided into nine training examples and one test example for each of the four classes. Training and testing were performed. This process was repeated 1000 times and the results averaged to yield a performance estimate of the system.

In the provided data set, all examples are time aligned precisely. Furthermore, the provided data set is assumed to have infinite SNR. In practice, neither condition can be assured. Performance of the class-dependent technique using the four proposed base representations, in the presence of noise and timing jitter, is investigated. Four cases presented here are as follows.

- Case 1—Classification using the original data that contains precise time alignment between all examples.
- Case 2—Classification using the original data with timing jitter uniformly distributed over the interval  $\pm$  one sample.
- Case 3—Classification using the original data with additive white Gaussian noise (AWGN) yielding an SNR of 14 dB.
- Case 4—Classification using the original data with timing jitter uniformly distributed over the interval  $\pm$  one sample and AWGN, yielding an SNR of 14 dB.

The performance of the class-dependent approach using four base representations (Rihaczek, Wigner-Ville, and two different spectrograms) is explored. For all cases and all representations, a 180-by-180 point auto-ambiguity function was used. A correlation threshold of 0.95 was used to determine the points to exclude using FDR. Two different window lengths were used to compute the spectrogram: one length 10 (which we call the wideband spectrogram) and the other length 120 (which we call the narrowband spectrogram). It is of interest to note that neither the narrowband nor the wideband spectrograms are regular TFRs. The results of each experiment are given in Table III.

*Case 1—Ideal Data Set:* All radar pulses in the original data set are perfectly time aligned with respect to the envelope of the signal. Using our Fisher's discriminant ratio kernel, no er-

rors were made on the test set with any of the three base representations. While we expect the spectrogram performance to be inferior to the Rihaczek and Wigner-Ville (for reasons detailed earlier), under these idealized conditions, performance is the same for both the narrowband and wideband versions.

*Case 2—Timing Jitter:* Timing jitter uniformly distributed over the interval  $\pm$  one sample was introduced. With the inclusion of jitter, an overall error rate of 0.0008 was achieved using the Rihaczek and Wigner-Ville TFRs. This represents three errors out of 4000. The narrowband spectrogram performed exactly the same as the Rihaczek and Wigner-Ville TFRs. The wideband spectrogram performed slightly worse with 17 errors out of 4000, corresponding to an overall error rate of 0.0043. The performance of the Rihaczek and Wigner-Ville TFRs demonstrate the ability of our approach to cope with misalignment in the input waveform adaptively through training. The advantage of the class-dependent approach is that from the same auto-ambiguity function projection our technique is able to design classifiers that are either sensitive or insensitive to time alignment *automatically*, depending on what is required for the particular scenario.

*Case 3—Noise:* The initial data set is idealized. AWGN was added to the original data lowering the SNR to 14 dB. This case is intended to mimic a more realistic environment. Under these conditions, 198 errors out of 4000 test examples were made using the Rihaczek TFR, corresponding to an error rate of 0.0495. The Wigner-Ville and narrowband spectrogram achieved the same performance as the Rihaczek TFR. The wideband spectrogram performance degraded significantly with 360 errors on the test set corresponding to an error rate of 0.09. On this data set, we conclude that the class-dependent approach is more susceptible to noise than timing jitter.

*Case 4—Timing Jitter and Noise:* This scenario combines the effects of noise and timing jitter. Timing jitter uniformly distributed over the interval  $\pm$  one sample along with AWGN to reduce the SNR to 14 dB was introduced into the data set. With this degradation of the input signal, performance dropped to 304 errors in the test set for the Rihaczek, Wigner-Ville, and narrowband spectrogram. This corresponds to an error rate 0.076. An example of the class-dependent TFR from each class is given in Fig. 6. Again, the performance of the spectrogram is inferior, with 476 errors on the test set corresponding to an error rate of 0.119. The difference in performance is a direct result of the zeros in the wideband spectrogram characteristic function, as shown in Fig. 7. Notice that the narrowband spectrogram experiences no performance degradation with respect to the Rihaczek/Wigner-Ville distributions. While there are zeros in the narrowband spectrogram characteristic function, the zeros are not located at points used for classification.

This result is intuitive; as we see from Fig. 7(c), the optimal points are located near the  $\eta = 0$  axis, corresponding to relatively stationary information. It is not surprising that the narrowband spectrogram (which has high frequency resolution and low time resolution) would surpass the performance of the wideband version (which has low frequency resolution and high time resolution). It is important to remember that by using either the narrowband or wideband versions of the spectrogram, implicit assumptions are made about the underlying time-frequency struc-

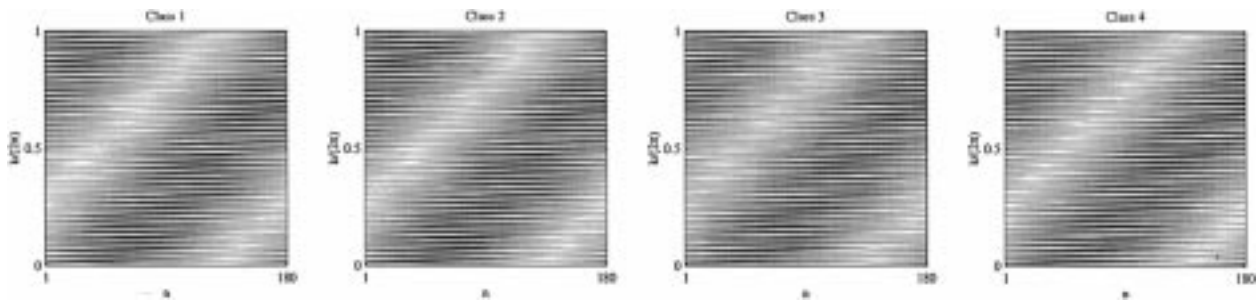


Fig. 6. Example of the log magnitude of the class-dependent TFR derived from the Rihaczek base TFR for each class under Case 4. These are generated using three kernel points. The largest magnitude is represented by the darkest gray-scale value.

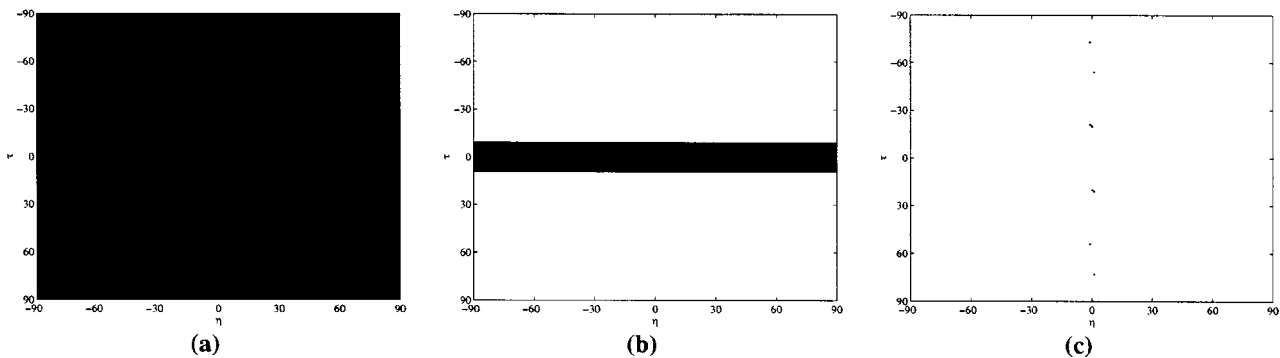


Fig. 7. (a) Dark areas indicate the nonzero points in the Rihaczek and Wigner–Ville transformation kernel. The Rihaczek is all “1,” and the Wigner–Ville is a complex modulating exponential. (b) Dark areas approximately indicate the nonzero points in the wideband spectrogram transformation kernel with a rectangular window of length 10. Notice that this representation is not regular. (c) Class-dependent kernel derived from the Rihaczek TFR for Case 4. These were used to achieve 92.4% correct classification. The optimal points are not available (i.e., they are set to zero) in the wideband spectrogram characteristic function. Therefore, with the wideband spectrogram base representation, other suboptimal points must be used, resulting in the performance loss observed in Case 4 (88.1%).

ture that differentiates classes (i.e., whether time or frequency resolution is important). This assumption is manifested as zeros in the characteristic function of the representation. By using the Rihaczek or Wigner–Ville base distributions, no such assumptions are made. In summary, by using regular TFRs, no assumptions are made about what time–frequency structure is important for classification, and thus, the performance of regular TFRs will always surpass (or perform at least as well as) TFRs that are not regular (which make assumptions about the type of time–frequency information that is important).

These results are for single pulse classification. If multiple independent radar pulses were used, classification performance would improve. A simple approach would be to combine multiple pulses after the decision. Although this approach to combining is suboptimal [31], using the Rihaczek or Wigner–Ville TFR with five pulses, the overall error rate would decrease to 0.0039 for Case 4 (assuming that out of five radar pulses, correctly classifying three, four, or five results in an overall correct classification). In typical radar applications, obtaining multiple pulses is a trivial matter due to the transmit rate.

Finally, we compare the class-dependent approach with some simple alternative feature sets. Four distinct sets were selected as follows:

- the class-dependent features derived from a Rihaczek base representation (determined in the previous section);
- the set of 180 autocorrelation points;
- the set of 180 Fourier coefficients;

- a set of 256 discrete wavelet coefficients computed using the Daubechies 4 wavelet and zero padding the end of the radar signal with 76 points.

In each example, the features were rank ordered using FDR. Classification then proceeded as described earlier. The results of this experiment are given in Table IV.

*Case 1—Ideal Data Set:* Both the class-dependent and autocorrelation features were able to classify these examples perfectly. The Fourier and wavelet features appear to be ill suited to this problem.

*Case 2—Timing Jitter:* With the inclusion of jitter, an overall error rate of 0.0008 was achieved using the class-dependent features. This represents three errors out of 4000. The autocorrelation features performed the same as Case 1 due their invariance to time shifts. The Rihaczek features experience slight degradation in performance because some of the selected kernel points are not on the  $\eta = 0$  axis.

*Case 3—Noise:* Under these conditions, the autocorrelation features experience a significant degradation in performance, committing 383 errors out of 4000 compared with 198 errors using the class-dependent features. This is a 48.3% decrease in error rate using class-dependent features compared with autocorrelation features.

*Case 4—Timing Jitter and Noise:* With this degradation of the input signal, performance dropped to 304 errors in the test set for the class-dependent features. The performance of the autocorrelation features remained the same as Case 3 due to their invariance to time shifts. This is a 20.7% decrease in the error

TABLE IV  
ERROR RATE OF THE CLASS-DEPENDENT APPROACH USING A NUMBER OF  
FEATURE SETS ON THE RADAR TRANSMITTER DATA. ONE THOUSAND  
EXAMPLES PER CLASS WERE USED TO ESTIMATE CLASSIFIER PERFORMANCE

Case	Feature Set			
	Class- Dependent	Autocorrelation	Fourier	Wavelet (derived from Daubechies 4)
1	0	0	0.5813	0.2438
2	0.0008	0	0.5917	0.4795
3	0.0495	0.0958	0.6927	0.6793
4	0.0760	0.0958	0.6895	0.7045

rate using class-dependent features compared with autocorrelation features.

A brief analysis of the computational demands of the class-dependent and autocorrelation features is provided. The computation will be expressed in terms of the number of complex multiplies required to compute the set of features. Each autocorrelation point requires  $N$  complex multiplies. Generation of the entire set of  $N$  autocorrelation points for a single training example requires  $N^2$  multiplies. Each auto-ambiguity point (class-dependent feature) requires  $2N$  complex multiplies. To generate the entire set of  $N^2$  auto-ambiguity points, for a single training example, requires  $2N^3$  multiplies. The computational requirements of the auto-ambiguity features are significantly higher than the autocorrelation features *during training*. However, the class-dependent approach compares favorably when performing actual classification (i.e., implementation of a previously trained system for real-time classification). This occurs because of the small number of auto-ambiguity points required for classification. For example, in Case 4, both the autocorrelation and Rihaczek features require four points to achieve best classification performance. To compute autocorrelation features for Case 4 requires 720 complex multiplies ( $180 \times 4$ ). Class-dependent features require only twice this value ( $(2 \times 180) \times 4$ ) to achieve the 20.7% decrease in error rate. Thus, to implement a previously trained class-dependent system, a relatively small increase in computational complexity is incurred to achieve a significant decrease in error rate. These figures represent a baseline algorithm; optimizations should be able to reduce the computational requirements presented here.

## VI. CONCLUSION

We have presented an approach to classification using time-frequency features. This approach makes no implicit assumptions about the amount and type of time-frequency smoothing required for classification. Making such assumptions may degrade classification. In general, any time-frequency classification technique that uses a singular TFR (e.g., the spectrogram, cone-kernel, or Choi-Williams) as a source of features will never surpass the performance of the same technique using a regular TFR (e.g., Rihaczek or Wigner-Ville). Use of singular TFRs implicitly discards information without explicitly determining if it is germane to the classification task. This has been empirically validated on the radar-transmitter identification task.

We propose to design and use the classifier directly in the auto-ambiguity plane. Since all TFRs can be derived from the auto-ambiguity plane, no *a priori* assumptions are being made about the smoothing required for accurate classification. Two techniques, whereby these kernels can be estimated, are presented. Training data is used to explicitly determine the kernels and to implicitly determine the amount and type of time and frequency resolution required for accurate classification. As an added advantage, our approach allows joint optimization of kernels from multiple simultaneous representations. This is advantageous when multiple sensors capture information on the same event, as was demonstrated on the helicopter fault diagnosis task.

The simulated study shows that our approach compares favorably with other techniques that have been benchmarked on this data set, approaching the performance of the Bayes optimal classifier. The real-world pilot studies indicate that our approach merits further investigation for possible implementation in an on-line system. It has been shown that with a small increase in computational complexity, a significant decrease in error rate is achieved.

A frequently encountered problem in the design of class-dependent kernels is the small number of examples relative to the number of parameters to be estimated for the kernel. This makes standard linear or quadratic discriminant functions problematic. We avoid this problem by first ranking with FDR (to reduce dimensionality) and then using a quadratic discriminant function. Other approaches may help alleviate this problem without ignoring the correlation that exists between points in the auto-ambiguity plane.

Finally, we mention that these techniques have recently been applied to tool-wear monitoring with interesting preliminary results [32]. An extension of this approach to unsupervised classification is also explored in that paper.

## ACKNOWLEDGMENT

The authors would like to thank Dr. J. Droppo for his comments and suggestions during this work. They would also like to thank Dr. V. Chen of the Naval Research Laboratory and Dr. A. Garga of the Penn State Applied Research Lab for providing the radar transmitter and helicopter transmission data, respectively.

## REFERENCES

- [1] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [2] W. J. Wang and P. D. McFadden, "Application of orthogonal wavelets to early gear damage detection," *Mech. Syst. Signal Process.*, vol. 9, no. 5, pp. 497–507, 1995.
- [3] P. M. Bentley, P. M. Grant, and J. T. E. McDonnell, "Time-frequency and time-scale techniques for the classification of native and bioprosthetic heart valve sounds," *IEEE Trans. Biomed. Eng.*, vol. 45, pp. 125–128, Jan. 1998.
- [4] Q. 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," *IEEE Trans. Signal Processing*, vol. 46, pp. 1202–1207, May 1998.
- [5] C. Heitz, "Optimum time-frequency representations for the classification and detection of signals," *Appl. Signal Process.*, vol. 2, no. 3, pp. 124–143, 1995.
- [6] M. Davy and C. Doncarli, "Optimal kernels of time-frequency representations for signal classification," in *Proc. IEEE-SP Int. Symp. Time-Freq. Time-Scale Anal.*, 1998, pp. 581–584.

- [7] S. Haykin and T. K. Bhattacharya, "Modular learning strategy for signal detection in a nonstationary environment," *IEEE Trans. Signal Processing*, vol. 45, pp. 1619–1637, June 1997.
- [8] S. Narayanan, J. McLaughlin, and J. Droppo, "Operator theory approach to discrete time-frequency representations," in *Proc. IEEE-SP Int. Symp. Time-Freq. Time-Scale Anal.*, 1996, pp. 521–524.
- [9] J. McLaughlin, J. Droppo, and L. Atlas, "Class-dependent time-frequency distributions via operator theory," in *Proc. ICASSP*, vol. 3, 1997, pp. 2045–2048.
- [10] L. Atlas, J. Droppo, and J. McLaughlin, "Optimizing time-frequency distributions via operator theory," *Proc. SPIE*, vol. 3162, pp. 161–171, 1997.
- [11] J. McLaughlin, "Applications of operator theory to time-frequency analysis and classification," Ph.D. dissertation, Univ. Washington, Seattle, 1997.
- [12] B. W. Gillespie and L. E. Atlas, "Optimization of time and frequency resolution for radar transmitter identification," *Proc. SPIE*, vol. 3461, pp. 91–98, 1998.
- [13] J. Droppo and L. Atlas, "Applications of classifier-optimal time-frequency distributions to speech analysis," in *Proc. IEEE-SP Int. Symp. Time-Freq. Time-Scale Anal.*, 1998, pp. 585–588.
- [14] A. W. Rihaczek, "Signal energy distributions in time and frequency," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 369–374, 1968.
- [15] T. A. C. Claasen and W. F. G. Mecklenbräuer, "The Wigner distribution—A tool for time-frequency signal analysis—Part III: Relations with other time frequency signal transformations," *Philips J. Res.*, vol. 35, no. 6, pp. 372–389, 1980.
- [16] F. Hlawatsch, "Regularity and unitarity of bilinear time-frequency signal representations," *IEEE Trans. Inform. Theory*, vol. 38, pp. 82–94, Jan. 1992.
- [17] M. S. Richman, T. W. Parks, and R. G. Shenoy, "Discrete-time, discrete-frequency, time-frequency analysis," *IEEE Trans. Signal Processing*, vol. 46, pp. 1517–1527, June 1998.
- [18] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
- [19] J. B. Buckheit and D. L. Donoho, "Improved linear discrimination using time-frequency dictionaries," *Wavelet Applications in Signal and Image Processing III—Proc. SPIE*, vol. 2569, no. 2, pp. 540–541, 1995.
- [20] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
- [21] D. H. Kil and F. B. Shin, *Pattern Recognition and Prediction with Applications to Signal Characterization*. Woodbury, NY: AIP, 1996.
- [22] G. H. Golub and C. E. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.
- [23] L. Breiman, J. H. Freidman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
- [24] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *J. Amer. Stat. Assoc.*, vol. 89, no. 428, pp. 1255–1270, 1994.
- [25] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *J. R. Stat. Soc. B—Methodol.*, vol. 58, no. 1, pp. 155–176, 1996.
- [26] P. D. McFadden and J. D. Smith, "A signal processing technique for detection local defections in a gear from the signal average of the vibration," *Proc. Inst. Mech. Eng.*, vol. 199, no. C4, pp. 287–292, 1985.
- [27] W. D. Mark, "Analysis of the vibratory excitation of gear systems: Basic theory," *J. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1409–1430, 1978.
- [28] W. J. Wang and P. D. McFadden, "Early detection of gear failure by vibration analysis—I and II," *Mech. Syst. Signal Process.*, vol. 9, no. 5, pp. 193–215, 1995.
- [29] B. G. Cameron, "The Westland helicopter report," <http://wisdom.arl.psu.edu/Westland/welcome.htm>.
- [30] V. C. Chen, "Time-frequency/time-scale analysis for navy radar applications," <http://airborne.nrl.navy.mil/vchen/tftsa.html>.
- [31] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [32] B. W. Gillespie and L. E. Atlas, "Data-driven time-frequency classification techniques applied to tool-wear monitoring," in *Proc. ICASSP*, 2000.



**Bradford W. Gillespie** (S'97) received the B.S.E.E. degree from Clarkson University, Potsdam, NY, and the M.S.E.E. degree from the University of New Hampshire, Durham, in 1992 and 1994, respectively. He is currently pursuing the Ph.D. degree with the Electrical Engineering Department, University of Washington, Seattle.

During the summers of 1999 and 2000, he was an Intern at Microsoft Research, Redmond, WA, in the Signal Processing Group. Prior to attending the University of Washington, he was an Algorithm Development Engineer for Sanders, a Lockheed–Martin Company, Manchester, NH. His interests include audio signal processing, speech enhancement, human audio perception, time–frequency analysis, and pattern recognition.

Mr. Gillespie has been a recipient of the Microsoft Research Graduate Fellowship, the Intel Foundation Graduate Fellowship, and the ARCS Fellowship



**Les E. Atlas** (M'82) received the B.S.E.E. degree from the University of Wisconsin, Madison, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1979 and 1983, respectively.

In 1984, he joined the Electrical Engineering Department, University of Washington, Seattle, where he is currently Professor and Associate Chair for Research. He co-founded the University of Washington's Interactive Systems Design Laboratory, where he is investigating time–frequency and other signal processing representations for machine learning in acoustical, mechanical, and manufacturing sensor signal processing applications. He also has research interests in auditory processing, speech processing, high-quality music coding, and statistical signal processing.

Dr. Atlas received the National Science Foundation's Presidential Young Investigator Award in 1985. He was General Chair of the 1992 IEEE International Symposium on Time–Frequency and Time-Scale Analysis and General Chair of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing. He is a Member-at-Large of the IEEE Signal Processing Society's Board of Governors and is Chair of the Society's Technical Committee on Signal Processing Theory and Methods.