Signal Processing for Multimedia
J.S. Byrnes (Ed.)
IOS Press, 1999

Optimizing Time and Frequency Resolution for Detection and Classification

Les Atlas

University of Washington

ABSTRACT. Research in time-frequency representations (TFRs) has often been directed towards determining how two-dimensional weighting kernels, which operate convolutionally on Wigner-Ville distributions, effect desired properties and trade-offs of the resulting representation. For example, a kernel with a diamond-shaped support region results in a spectrogram which has the well-known trade-off between time and frequency resolution. Much past research has been directed at improving resolution, while ameliorating the quadratic interference of the Wigner-Ville approach. We take an entirely different view: Our final goal is data-trained pattern classification, where high resolution may only increase the need for training data. We thus change the standard approach to automatically determine the kernel which minimize the time and frequency resolution needed to differentiate multiple classes. The kernels are called "class-dependent kernels." We have applied these class-dependent kernels to problems in multi-sensor helicopter fault diagnosis. In this application, perfect detection of the occurance of a fault and perfect classification of the type of fault was achieved. Also, the optimal sensors for each fault were automatically chosen.

1. Introduction

Vibration patterns from accelerometers mounted on gearboxes are, in general, analyzed by using time- or frequency-domain information. Past researchers have applied either time-domain (e.g. [1]), frequency-domain (e.g. [2]), or combined time-frequency (e.g. [3]) approaches to fault diagnoses in gearboxes. In the latter case, the application of time-frequency or related (e.g. [4]) techniques have used transform-limited or empirically-determined time and frequency resolutions for representations of examples of healthy and failed gearboxes. Our recent approach has been to instead allow the classification or detection task to, given adequate and representative training data, optimally determine the relative role of time and frequency resolution in the representation [5,6]. For example, if transients occurring at a certain phase of rotation are important for detecting imminent failure due to crack formation, it would be best to have high resolution in time. Alternatively, if a harmonic or a set of harmonics at certain regions in frequency are important for detecting root fatigue, it would be best to have high resolution within these frequency regions.

Our approach is also based on the premise that automatic detection or classification systems should be provided with only enough input resolution to achieve needed performance. Namely, too great a resolution will potentially require much too large of a detector or classifier training data set and will also be sensitive to irrelevant features and/ or noise. Also, large dimensionality detectors or classifiers are computationally expensive and slow. It should also be noted that we are not referring to or bound by implicit Heisenberg or window-related resolution limitations—we are instead explicitly limiting the resolution to optimize the best early detection or classification of faults.

2. Background

Modern time-frequency representation (TFR) research often begins with choosing a kernel function $\Phi[n,m]$ which operates upon an instantaneous autocorrelation function

$$\Upsilon[n,m] \equiv \sum_{n'=n}^{N} x[n']x[n'-m].$$

The resultant TFR ($P[n,k]$) arises from the discrete Fourier transform (in m) of the result of multiplying the kernel (in m) and convolving the kernel (in n) with this instantaneous autocorrelation function ($\Upsilon[n,m]$). As an alternative, a discrete Fourier transform (in n) can be applied to this instantaneous autocorrelation function ($\Upsilon[n,m]$) to provide an ambiguity function

$$A[l,m] = \mathcal{F}_n\{\Upsilon[n,m]\} \equiv \sum_{n=0}^{M-1} \Upsilon[n,m]e^{-j\frac{2\pi}{M}nl}.$$

There is an equivalent kernel ($\phi[l,m]$) which operates multiplicatively in both dimensions upon this ambiguity function ($A[l,m]$). These two kernels are also related by a discrete Fourier transform (in n)

$$\phi[l,m] = \mathcal{F}_n\{\Phi[n,m]\} \equiv \sum_{n=0}^{M-1} \Phi[n,m]e^{-j\frac{2\pi}{M}nl}.$$

Any non-zero extent of $\phi[l,m]$ in l and/or m can effect a smoothing of $P[n,k]$ in time and/or frequency, respectively, creating a new "smoothed" distribution $CC[n,k]$. For example, if $\phi[l,m] = 0$ for all values except for those on the $l = 0$ axis, then all temporal information is smoothed and only steady-state frequency information in $P[n,k]$ is retained in $CC[n,k]$. In past time-frequency research, kernels for quite a number of properties, such as finite-time support and minimizing quadratic interference, have been determined. Though some of these past time-frequency representations may offer advantages in classification of certain types of signals, the goal of sensitive detection or accurate classification has not been explicit. The above kernel's ($\phi[l,m]$) capacity to reduce time and/or frequency resolution, embodied within the explicit goal of optimal classification (i.e. minimum number of classification errors), is the basis of the approach we outline below. When $P[n,k]$ is smoothed with this goal the resulting TFR, $CC[n,k]$, is called the "class-conditional TFR."

3. Our Approach and Methods

We utilized data provided by the Applied Research Laboratory (ARL) at Penn State. This data contained eight separate and individual fault types and associated normal (no fault) conditions. These data were collected utilizing Westland Helicopters Ltd.'s Universal Transmission Test Rig to test a CH46 aft transmission. Eight accelerometer time series were produced for each fault condition at varying torque levels [8].

The multi-dimensional time series for each fault was divided into 1500-point, nonoverlapping and contiguous one-dimensional segments ($x_{f,a,i}[n]$) where f was the type of fault ($f = 9$ for no fault), a was the accelerometer channel, and i was the index of the segment. (These same subscripts will also be used for subsequent kernel and ambiguity function notation.) Each segment was individually demeaned and scaled by its estimated standard deviation. The total number of segments for each fault was 274, where each was 14.5 milliseconds long.

In order to experimentally study a proposed fault detection/ classification system, the above data segments were randomly divided into equal-sized and non-overlapping training and test sets. The training sets were used to determine the parameters of the system below and the independent test sets were used to predict the performance.

Our approach is a modification of the signal class-dependent kernel approach which has been described in more detail before [5,6,7]. Simply stated, this approach finds the single kernel ($\phi_{fg,a}[l,m]$) which optimizes the mean-square distance between estimated ambiguity functions ($A_{f,a,i}[l,m]$ and $A_{g,a,i}[l,m]$) representing two different classes of data (f and g). The points in this kernel are then ranked according to separation between classes, where choosing the kernel point with the largest interclass separation corresponds to a maximum separation (in theory) between classes. Thus for actual classification of unknown time series, the ambiguity functions are multiplied (in l and m) by a binary kernel mask, which is set to "1" at one optimal, and optionally, subsequently lower-ranked kernel points (often required in practice). These kernel points, depending upon their locations, effect a smoothing in time and/or frequency of the unknown data. The smoothed version is then compared to smoothed representatives of training classes. As an added result, the kernels for optimal separation can be transformed to a time-frequency representation (the class-conditional TFRs $CC_{f,a}[n,k]$ and $CC_{g,a}[n,k]$) and the implicit optimal time-frequency smoothing can be viewed after training.

As we have most recently found for this Westland Data, the above mean-square distance is inadequate for handling the wide range of within-class variances seen. Thus, we have modified our earlier approach [5,6] to find the kernel $\phi_{fg,a}[l,m]$ which optimizes a Fisher's discriminant distance between estimated ambiguity functions $A_{f,a,i}[l,m]$ and $A_{g,a,i}[l,m]$, representing two different classes of data. The Fisher's discriminant for two classes, class f and class g, is represented by

$$FDR_a[l,m] \equiv \frac{|\frac{1}{I}\sum_i A_{f,a,i}[l,m] - \frac{1}{I}\sum_i A_{g,a,i}[l,m]|^2}{\frac{1}{I}\sum_i |A_{f,a,i}[l,m]|^2 - \frac{1}{I}\sum_i |A_{g,a,i}[l,m]|^2}$$

where the above sums index through the l segments of training data available for each class. The Fisher's discriminant distance provides a rank-ordering of kernel points for classification. The optimal number of points will be determined by evaluating the classifier performance using the K best (i.e. K points with the largest Fisher's discriminant distance) kernel points.

To classify a particular unknown test example, an M by M (in this paper we present data using either $M = 16$ or $M = 64$) ambiguity function was estimated for each 1500-point segment of the unknown time-series. After masking with the appropriate kernel, the unknown class is estimated via a Maximum Likelihood (ML) detector. The mean and covariance matrices (utilized by the ML detector) for each class are estimated from the training data. For our results reported below, we restricted our study to two classes: fault or no-fault discrimination. However, as outlined in [7], extensions to multiple classes are straightforward.

We also propose a measure of confidence in our classification. By finding a ratio R of the distance to the closest class over the distance to the other class, we can define a confidence measure:

$$Confidence \equiv \frac{1}{1+R}$$

which ranges from 0.5 to 1.0.

4. Results

We summarize our results with four key points and a few representative example plots.

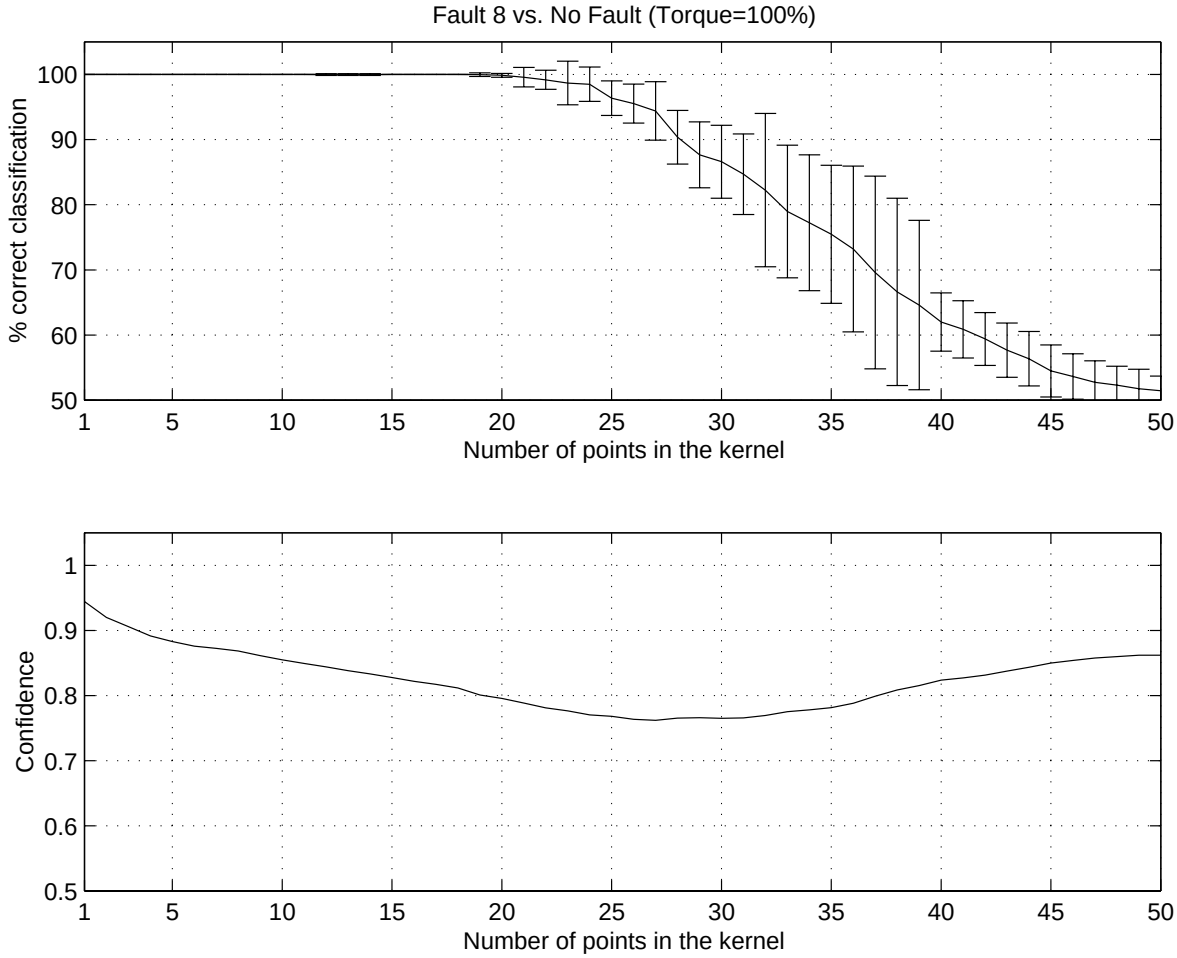

FIGURE 1. Classifier performance (top) and certainty (bottom) for discrimination between fault 8 (quill shaft crack propagation, run: #87 torque: 100%) and no fault (run: #01, torque: 100%) using accelerometer 4. The top figure shows the average performance plus or minus one standard deviation.

1. When the system is trained and then applied at consistent torque levels, it is possible to individually classify all eight faults in the Westland Data with 100by 16 ambiguity function and data from one accelerometer. Namely, by using the previously outlined approach to train one class with no-fault data at a given torque level and another class with data from one of the eight faults at the same torque level, the accuracy for automatic classification of left out (not in the training data) fault/no-fault conditions was 100100
2. The best point of the optimal kernel was on the kernels' ($\phi[l, m]$) $l = 0$ axes for all fault conditions. What this means is that only frequency and not time information was needed for the above 100given by Figure 2:

   A more illustrative representation of the same information would be the magnitude and phase of the difference between the fault and no-fault distributions. Figure 3 shows this difference for the class-conditional TFRs given in Figure 2. It is important to note that only magnitude and not phase information is important in the optimal classifier.
3. By including more kernel points, the classification performance drops for the single torque case. Also, a significant number of non-zero kernel points are now well off the kernel's l=0 axes and temporal information is included in the classification. Figure 4 shows an example result when the 50 kernel points were included:

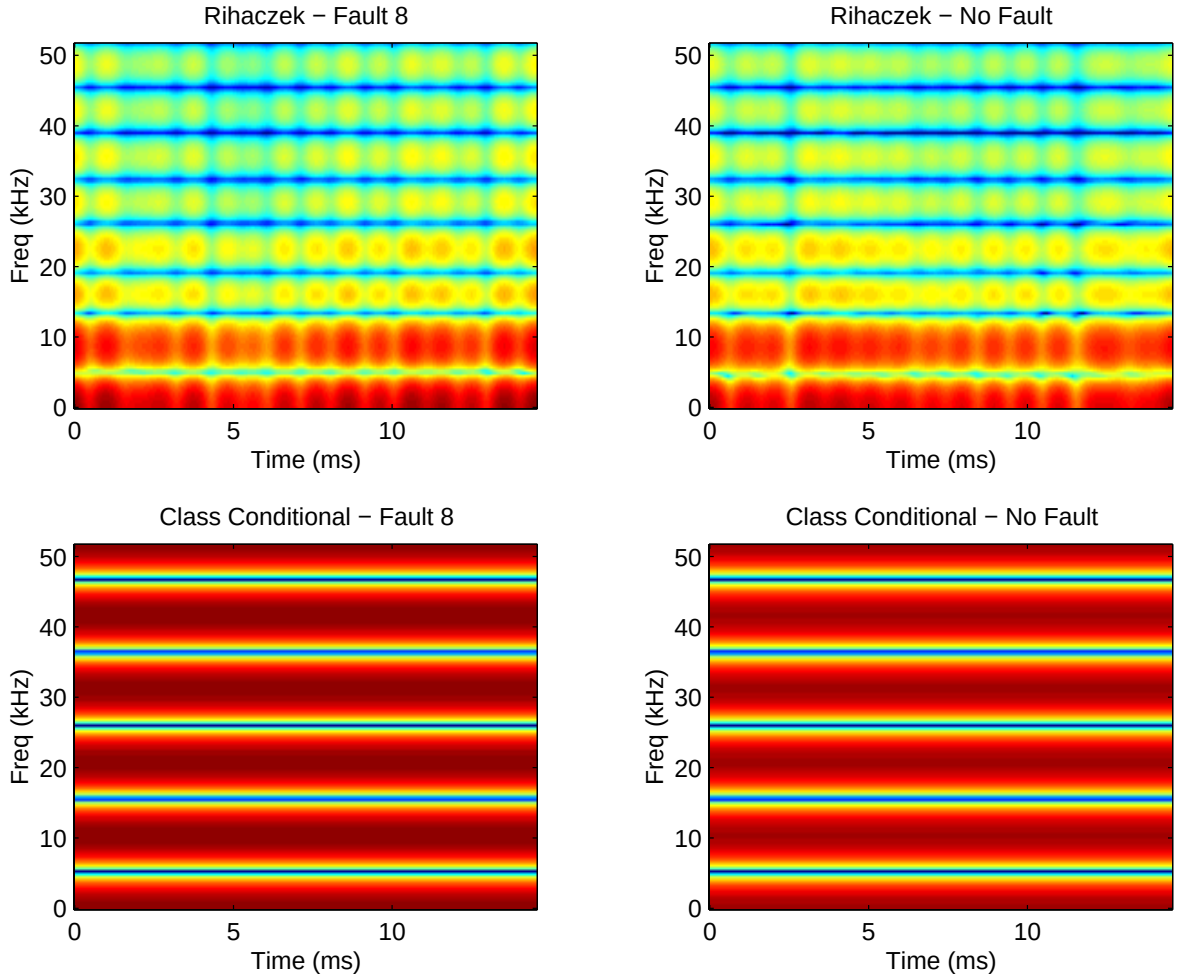

FIGURE 2. Magnitude of the unsmoothed Rihaczek time-frequency distributions (top) for fault 8 (quill shaft crack propagation, run: #87 torque: 100%) and no fault (run: #01, torque: 100%). Magnitude of the optimally smoothed class-conditional distributions for fault 8 and no fault classes (bottom) generated using 1 kernel point from accelerometer 4.

4. Classifier performance drops from the 100system is trained and tested with multiple torques using a 16 by 16 kernel. However, if the resolution of the initial time-frequency distribution is increased (by using a 64 by 64 kernel) we again achieve 100Figure 5 shows the performance of the classifier trained on four different torque levels using an initial TFR with a 64 by 64 point kernel. Optimal classification is achieved with 6 kernel points for this fault type. Also, the scaled, multiple-torque, case still only requires frequency information for optimal classification. Figure 6 shows the average magnitude and phase difference between fault and no fault class-conditional distributions

5. Summary of Performance

We propose two potential system configurations, one which requires torque information to classify faults and the other which does not. The system that did not require torque information (the simpler) achieved almost perfect classification over a variety of faults. The system that required a classifier for each fault and torque level tested achieved perfect classification under all conditions. The results of our analysis are presented below.

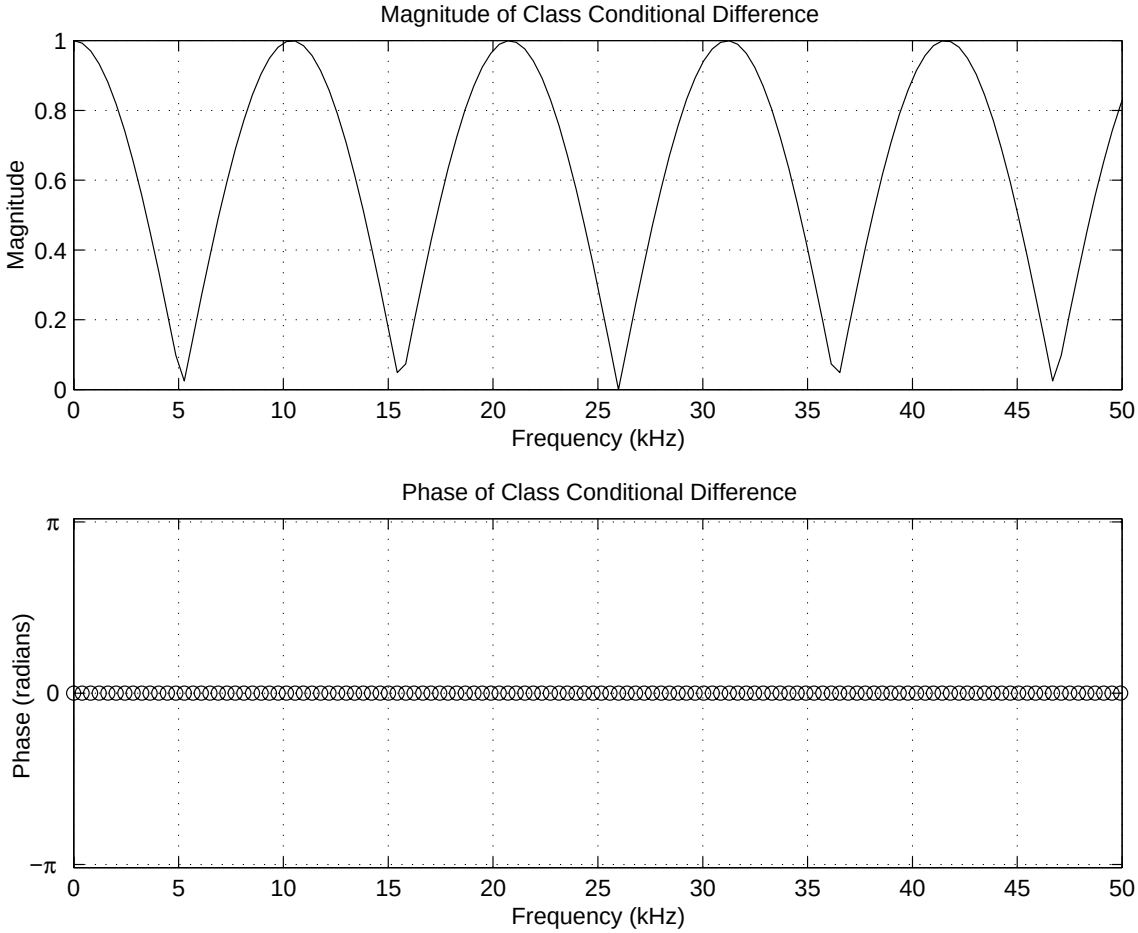

FIGURE 3. Differences between class-conditional magnitudes (top) and phases (bottom) for fault 8 (quill shaft crack propagation, run: #87 torque: 100%) and no fault (run: #01, torque: 100%) generated using 1 kernel point from accelerometer 4.

6. Conclusions

We have verified that our class-conditional time frequency approach can be used to automatically detect seeded faults with 100% accuracy. Implicit in these studies was the observation that frequency information alone, e.g. long-term spectral estimates, would have been adequate for this accuracy. We have also shown that constructing a time-frequency representation which is optimized for classification ab initio could be more efficient than applying classifiers and detectors to standard or high-resolution time-frequency representations. We have also found that an increase in resolution of the initial time-frequency representations is necessary to handle widely varying torque conditions. However, this multiple-torque system still requires frequency information alone. Further study and open discussion is needed to determine how our automatically determined stationary frequency weightings correspond with the mechanics and physics of the faults.

References

[1] P.D. McFadden and J.D. Smith, "A signal processing technique for detecting local defects in a gear from the signal average of the vibration," *Proc. Instn. Mech. Engrs.*, Vol. 199, No. C4, 1985, pp. 287–292.

[2] W.D. Mark, "Analysis of the vibratory excitation of gear systems: Basic theory," *J. Acous. Soc. Am.*, Vol. 63, No. 5, May 1978, pp. 1409–1430.

[3] W.J. Wang and P.D. McFadden, "Early detection of gear failure by vibration analysis—I. and II.," *Mechanical Systems and Signal Processing*, Vol. 7, No. 3, 1993, pp. 193–215.

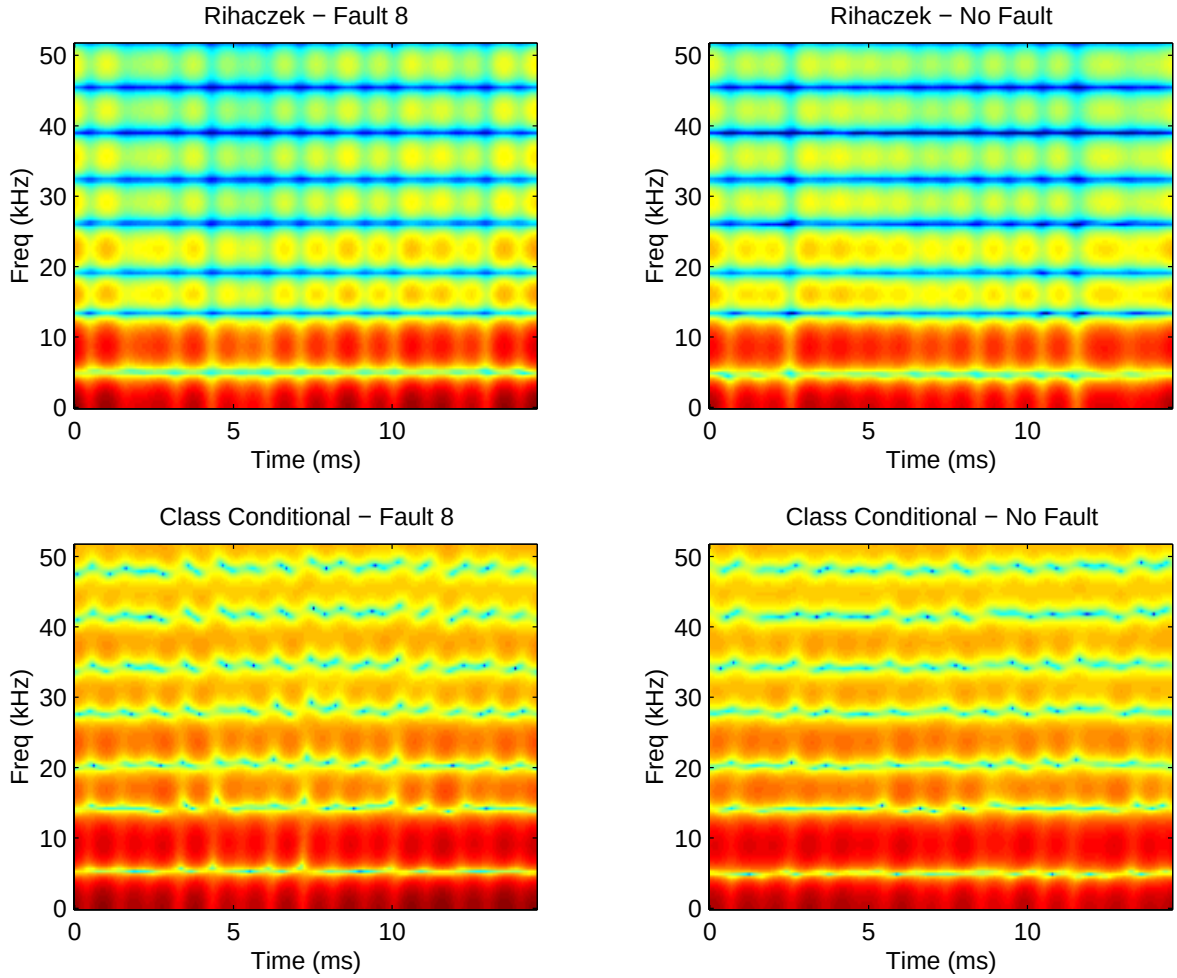

FIGURE 4. Magnitude of the unsmoothed Rihaczek time-frequency distributions (top) for fault 8 (quill shaft crack propagation, run: #87 torque: 100%) and no fault (run: #01, torque: 100%). Magnitude of the suboptimally smoothed 50-point class-conditional distributions for fault 8 and no fault classes (bottom) generated using 50 kernel points from accelerometer 4.

[4] W.J. Wang and P.D. McFadden, "Application of orthogonal wavelets to early gear damage detection," *Mechanical Systems and Signal Processing*, Vol. 9, No. 5, 1995, pp. 497–507.

[5] J. McLaughlin, J. Droppo, and L. Atlas, "Class-dependent time-frequency distributions via operator theory," in *Proceedings of the 1997 IEEE ICASSP*, vol. 3, (Munich, Germany), 1997, pp. 2045–2048.

[6] L. Atlas, J. Droppo, and J. McLaughlin, "Optimizing time-frequency distributions for automatic classification," in *Proceedings of the SPIE-The International Society for Optical Engineering*, Vol. 3162 (San Diego, CA), 1997.

[7] J. McLaughlin and L. Atlas, "Applications of operator theory to time-frequency analysis and classification," submitted to *IEEE Transactions on Signal Processing*, 1997

[8] B.G. Cameron, "The Westland Helicopter Report," `http://wisdom.arl.psu.edu/Westland/-welcome.htm`, Current as of April 15, 1998.

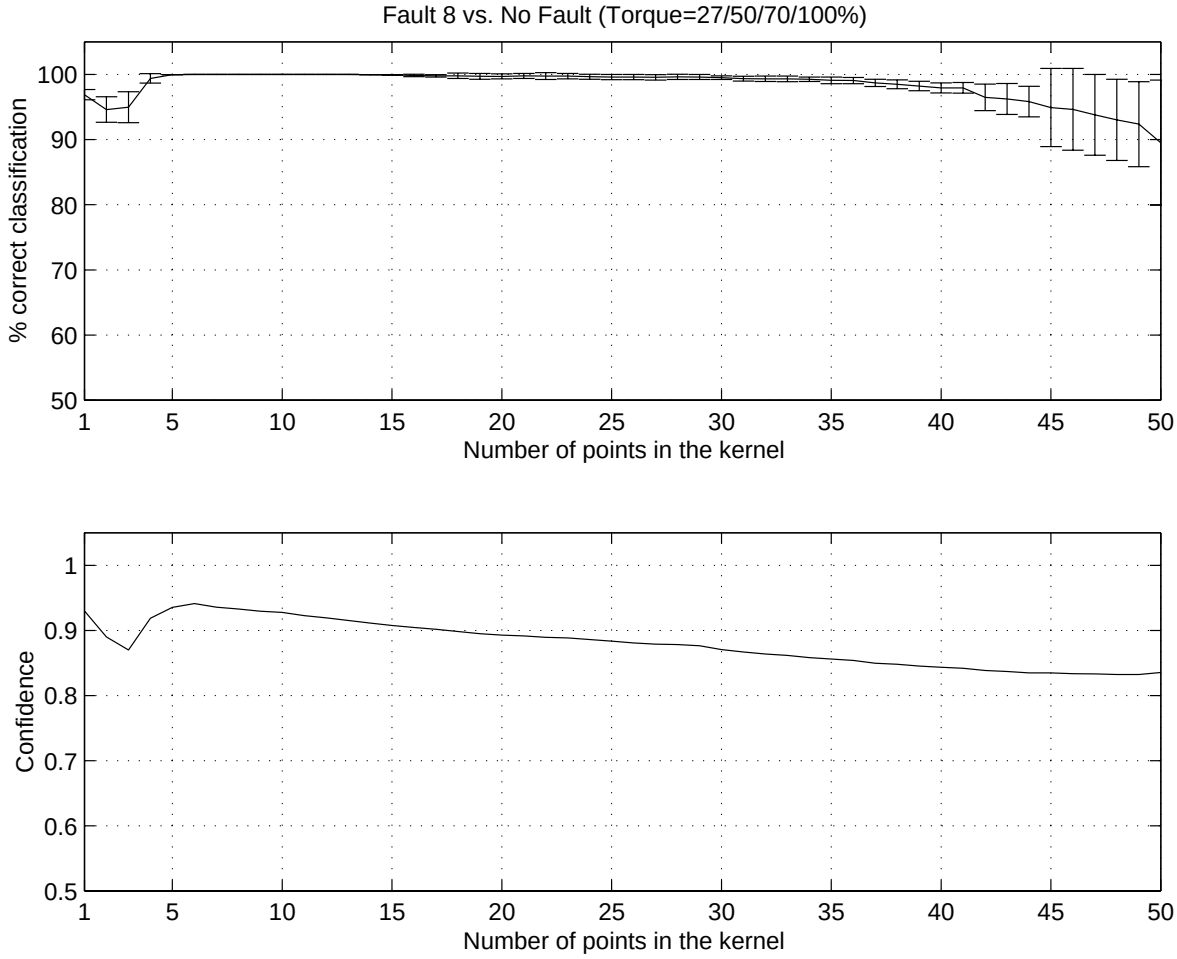

FIGURE 5. Classifier performance (top) and certainty (bottom) for discrimination between fault 8 (quill shaft crack propagation, run: #79 with torque: 27%, run: #82 with torque: 50%, run: #84 with torque: 70%, run: #87 with torque: 100%) and no fault (run: #02 with torque: 27%, run: #03 with torque: 50%, run: #04 with torque: 70%, run: #01 with torque: 100%) using accelerometer 4. The top figure shows the average performance plus and minus one standard deviation.

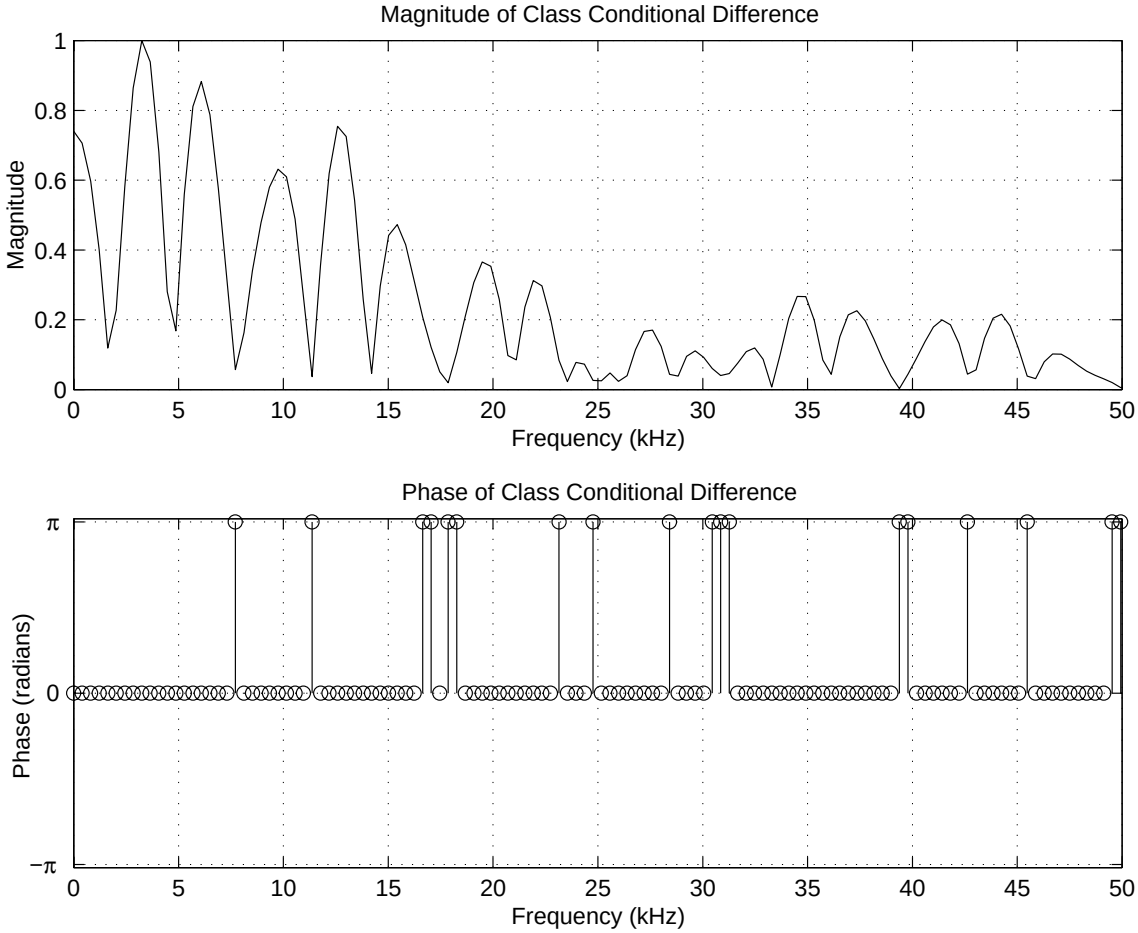

FIGURE 6. Differences between class-conditional magnitudes (top) and phases (bottom) for fault 8 (quill shaft crack propagation, run: #79 with torque: 27%, run: #82 with torque: 50%, run: #84 with torque: 70%, run: #87 with torque: 100%) and no fault (run: #02 with torque: 27%, run: #03 with torque: 50%, run: #04 with torque: 70%, run: #01 with torque: 100%) generated using 6 kernel points from accelerometer 4.

Fault Type	Best Accelerometer	Optimal Number of Kernel Points	Error Rate (% correct)
Epicyclic Planet Gear Bore/Bearing/ Inner Race Corrosion Spalling Defect (Fault 3)	3	21	99.8%
Spiral Bevel Input Pinion Spalling (Fault 4)	7	10	100%
Collector Gear Crack Propagation (Fault 7)	1	17	99.9%
Quill Shaft Crack Propagation (Fault 8)	4	6	100%

FIGURE 7. Classifier Performance over full range of torque levels (27, 50, 70 and 100% torque)

Fault Type	Best Accelerometer	Optimal Number of Kernel Points	Error Rate (% correct)
Epicyclic Planet Gear Bore/Bearing/ Inner Race Corrosion Spalling Defect (Fault 3)	3	Depends on torque	100%
Spiral Bevel Input Pinion Spalling (Fault 4)	7	1	100%
Collector Gear Crack Propagation (Fault 7)	1	1	100%
Quill Shaft Crack Propagation (Fault 8)	4	Depends on torque	100%

FIGURE 8. Classifier Performance for a Fixed Torque Level (27, 50, 70 or 100% torque)