A Multi-Fault Modeling Approach for Fault Diagnosis and Failure Prognosis of Engineering Systems

Bin Zhang[1], Chris Sconyers[2], Romano Patrick[1], George Vachtsevanos[2]

[1] *Impact Technologies LLC, Rochester, NY, 14623*
bin.zhang@impact-tek.edu; romano.patrick@impact-tek.com

[2] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332*
csconyers@gatech.edu; gjv@gatech.edu

ABSTRACT

Accurate and reliable fault diagnosis and prognosis of safety or mission critical components/ subsystems in complex engineering systems present major challenges to the Condition-Based Maintenance (CBM) or Prognostic and Health Management (PHM) designer. A crucial step in the development of CBM/PHM strategies relates to the designer's ability to understand and model the incipient failure or fault modes and mechanisms. A single fault growth model might not be often capable to capture a sequence of fault behaviors. Consider, for example, a rolling element bearing as a critical component of rotating machinery. The bearing may begin to corrode under certain operating conditions and, in parallel or sequentially, may be spalling and eventually, cracking. For accurate model-based fault diagnosis and failure prognosis, therefore, it is essential that fault progression models be developed to represent these evolving behaviors. This paper introduces an approach to multi-fault modeling with an application to a rolling element bearing of a helicopter's oil cooler. A simple and cost-effective on-line parameter adaptation solution is introduced to improve the performance of modeling. Finally, a series of experiments for different fault modes are presented to verify the proposed solution.

1 INTRODUCTION

Rotating machinery is widely used in various industrial, military, and commercial processes. Rolling element bearings are essential components in such applications and their failures often result in a critical damage, downtime, and costly repair [4]. Therefore, fault diagnosis and failure prognosis, which provide a condition based maintenance strategy to either machinery or components, such as bearings, is important to the safety of the system and results in substantial economic benefits [1-3, 10-12].

The bearing faults are usually closely related to speed, load, and operating environments (including grease condition). Compared to speed, the load and environmental conditions play an important role in the bearing health condition. When bearings are working in harsh environments, such as a humid and dusty environment, it is very likely that grease breakdown, grease wash-away, or debris denting can happen. If water or corrosive materials in the bearing grease reach a quantity such that the lubricant cannot protect the raceway surface, destructive chemical or electromechanical reaction of materials may result in rust or corrosion. Corrosion is often the first symptom of a bearing fault.

When corrosion develops, more debris is peeled off resulting in a rough surface of the raceway. At the same time, the debris may penetrate the grease resulting in further degradation of the greasing condition. This will cause a larger friction between the rolling elements, larger irregular vibration and higher temperature. Friction and high temperature can lead to softening of the bearing and a reduction of its load carrying capacity. As a result, corrosion finally develops into spalling.

In the process of spall development, loads, especially excessive or aggressive loads, are also leading contributing factors. Excessive loads usually cause fatigue failure due to repeated contact between the raceway and the rolling elements. Fatigue appears first below the surface because the contact between the raceway and the rolling elements creates stresses reaching a maximum value below the surface and slowly progressing to the surface. When they reach the surface, peeling or spalling or cracking of the surface of the raceway or rolling elements occurs. Corrosion and spalling can also initiate cracking and/or flaking of material. Cracking in a bearing is a severe fault mode.

It is evident that during the lifespan of a bearing, an incipient failure may manifest itself in different stages each one exhibiting different symptoms with the latter appearing with a variety of characteristic signatures in the on-line measurements. Evidence also supports the fact that different

fault stages may respond more strongly to particular sensing modalities. It is desirable, therefore, to detect the fault stages, to distinguish between them and to accurately predict the initiation and progression of each fault stage.

Paris' law is widely used in fatigue crack growth prediction. Based on Paris' law, a modified law is proposed in [21]. A relationship between the impulse magnitude and the spall size is given in [15] and an inverse model is built based on this relationship to track a bearing spall [22]. In [16], a hardness number is considered in the modeling module since it affects the crack progression [17]. Ioannides *et al* [18] introduced the effect of fatigue limit stress in the life prediction. Since the relationship between the time to failure, running time and stiffness can be established from damage mechanics, the natural frequency and the acceleration amplitude of a bearing component can be related to its failure time and a model is based on this concept [19]. An extension to these notions suitable for a variety of condition indicators is given in [20].

This paper is organized as follows: Section II presents an architecture for fault diagnosis and failure prognosis that is suitable for multi-fault modes. Section III discusses the details on how to build fault progression models for different fault modes. The parameter adaptation, along with the influence of operating and environmental conditions, is addressed. Experimental results for different fault modes are presented in Section IV to verify the proposed methodology. Finally, concluding remarks are given in Section V.

2 FAULT DIAGNOSIS AND FAILURE PROGNOSIS

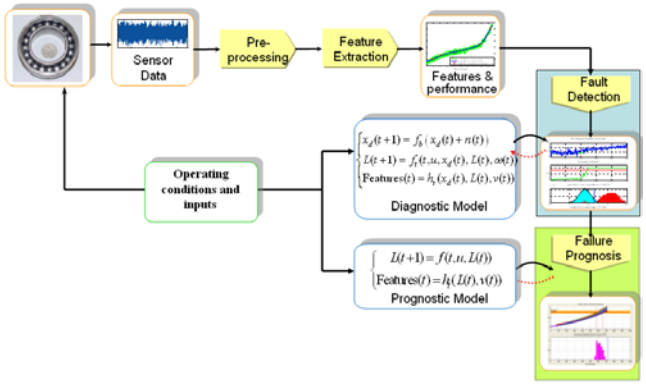

Figure 1. Proposed architecture for an anomaly detector

Figure 1 depicts the proposed fault diagnosis and failure prognosis architecture for a single fault mode. In this architecture, real-time measurements and operating conditions are provided in real time. Data are pre-processed before computing the features that will assist to efficiently monitor the behavior of the plant. With the features and a model describing the degrading state of the system, a fault detection algorithm based on particle filtering can be applied. Statistical analysis is implemented to arrive at the probability of a certain fault. When the fault is detected with a specified confidence level and given false alarm rate, a failure prognostic algorithm is activated to predict the Remaining Useful Life (RUL) of the component. This architecture provides not only a convenient compromise between data-driven and model-based techniques, but also the means to discuss its performance in terms of statistical performance indices. Moreover, a particle filtering-based algorithm enables us to efficiently deal with nonlinear systems and no-Gaussian noise.

In this architecture, the most important modules supporting the implementation of the algorithm are feature extraction and diagnosis/prognosis models. Features are the foundation for "good" fault detection algorithms. Usually, different fault modes show different fault characteristics in the raw measurements. Therefore, when more than one, say N, fault modes need to be detected, correspondingly N features need to be extracted. In this paper, we assume that for different fault modes, a series of features have been extracted regarding the fault modes of interest.

Our focus in this paper is the diagnosis and prognosis models. Before we proceed with the modeling details, it is desirable to extend the architecture, shown in Figure 1, so that it can be used to detect and predict the presence of more than one fault modes. For these different fault modes, as corrosion, spalling, and cracking of bearings, described in Section I, the former one usually leads to the next one.

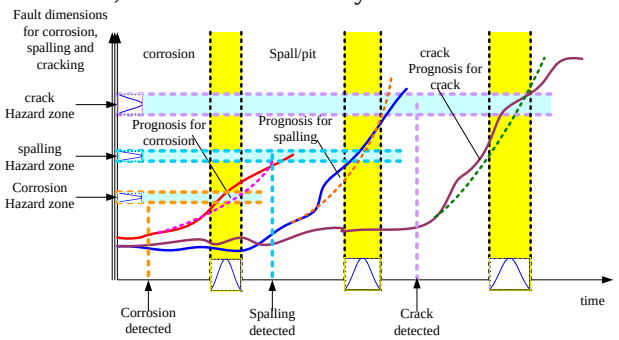

Figure 2. Multi-fault modes

When N fault modes are considered, due to the different fault mechanisms, we also need N models for diagnosis and prognosis purposes. Taking the bearing as an example, the proposed scheme is described as follows:

1. Build three models for fault progression of corrosion, spalling, and cracking. At this stage, from theoretical analysis and results from the published literature, model structures are built and parameters are initialized for further processing.
2. When new measurements (i.e. features for corrosion, spalling, and cracking) become available, the parameter sets for these different models are tuned on-line.
3. Three on-line diagnosis algorithms to detect corrosion, spalling and cracking are implemented in parallel. Since a typical fault detection routine is not computationally intensive, real-time processing of the three algorithms does not present major difficulties.
4. When a fault is detected, as shown in Figure 2, the prognosis routine for this fault mode is activated to predict the time evolution of this fault mode. Under the current situation, the detected fault mode is the dominant one. The prognostic routine predicts the

remaining useful life of the failing component or the time to reach a predefined threshold level, or hazard zone.

5. While the first prognostic routine is running and the next fault mode is detected, the corresponding prognostic algorithm is activated and the fault's evolution is tracked. This scheme is repeated for the third and final mode. The time progression of the three fault modes is tracked until one of them reaches its failure threshold or hazard zone. The latter is represented as a probability density function and estimated from historical failure data or experience. Depending on the prevailing usage patterns, operating and environmental conditions, anyone of the three fault modes may be expected to reach first its threshold value. For this reason, the prognostic routines are running in parallel although the dominance of one mode over the other two may be obvious from the evidence at hand and the interactive coupling between them.
6. Repeat step 5.

This way, a complete fault diagnosis and prognosis scheme is implemented for the component of interest. The benefit is a more accurate and precise prediction of the component's remaining life. In addition, the progression of faults provides a better understanding to the end-user of the component's anomalous behavior for correct mission and/or maintenance decisions.

3 DIAGNOSIS AND PROGNOSIS MODELS

Recent advances in model-based diagnostics and prognostics have focused on two fundamental approaches: The first one relies on physics of failure principles and determines the fault detection and time progression framework, as well as associated features or Condition Indicators, employing tools from Finite Element Analysis and a dynamic nonlinear description of the component/subsystem under faulty condition [6]. This approach facilitates an excellent understanding of component failure mechanisms and results in an optimum feature vector. Unfortunately, due to its complexity, it is time intensive and consumes extensive computational resources. An approach that alleviates these difficulties takes advantage of empirical or semi-empirical modeling tools, thus avoiding the complexity issues. The underpinnings of this second approach are found in population statistics, such as Paris' or Arrhenius' law and their variants. We adopt the second approach in our multi-fault modeling effort.

The fault progression is often nonlinear and, consequently, the model should be nonlinear as well. From a nonlinear Bayesian state estimation standpoint, diagnosis and prognosis may be accomplished by the use of a Particle Filter-based module [5]. An essential element of this module is a nonlinear state model describing the time progression or evolution of the fault.

3.1 Fault Diagnosis

A Fault Detection and Identification (FDI) procedure may be interpreted as the fusion and utilization of the information present in a feature vector (observations) with the objective of determining the operational condition (state) of a system and the causes for deviations from particularly desired behavioral patterns. Therefore, a model for diagnosis is given as follows:

$$\begin{cases} x_d(t+1) = f_b\left(x_d(t) + n(t)\right) \\ x_c(t+1) = f_t(x_d(t), x_c(t), \omega(t)) \\ \text{Features}(t) = h_t(x_d(t), x_c(t), v(t)) \end{cases} \quad (1)$$

where f_b, f_t and h_t are non-linear mappings, $x_d(t)$ is a collection of Boolean states associated with the presence of a particular operational condition (normal or faulty) in the system, x_c is the continuous-valued state that represents the fault dimension, $\omega(t)$ and $v(t)$ are non-Gaussian noises that characterize the process and feature noise signals, respectively, $n(t)$ is i.i.d. uniform white noise. Note that the last equation in (1) builds a connection between system state and feature vector.

Suppose that in diagnosis, only two states, normal and faulty, are under consideration. Moreover, assuming that the nonlinear mapping $h(.)$ between the feature and fault state is one-to-one, then, for an actual system, model (1) can be re-written to facilitate the implementation of the scheme shown in Figure 1:

$$\begin{bmatrix} x_{d,1}(t+1) \\ x_{d,2}(t+1) \end{bmatrix} = f_b\left(\begin{bmatrix} x_{d,1}(t) \\ x_{d,2}(t) \end{bmatrix} + n(t)\right)$$
$$x_c(t+1) = \left[(1+\beta)x_c(t)\right] \cdot x_{d,2}(t) + \omega(t) \quad (2)$$
$$y(t) = x_c(t) + v(t)$$
$$f_b(x) = \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix}^T, \text{ if } \left\| x - \begin{bmatrix} 1 & 0 \end{bmatrix}^T \right\| \le \left\| x - \begin{bmatrix} 0 & 1 \end{bmatrix}^T \right\| \\ \begin{bmatrix} 0 & 1 \end{bmatrix}^T, \text{ else} \end{cases}$$
$$\begin{bmatrix} x_{d,1}(0) & x_{d,2}(0) & x_c(0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

In this model, f_b is a non-linear mapping, $x_{d,1}$ and $x_{d,2}$ are Boolean states that indicate normal and faulty conditions, respectively, $y(t)$ is the noise-contaminated fault dimension, and β is a time-varying model parameter that describes the progression of the fault dimension under a fatigue stress.

3.2 Failure Prognosis

Prognosis is activated when a fault is detected. For the same fault mode, the progression of the fault follows the same physical law as the one governing the fault's detection. Therefore, in the prognosis model, the Boolean state of (1) can be replaced with the following expression:

$$\begin{cases} x_c(t+1) = f_t(x_c(t), \omega(t)) \\ \text{Features}(t) = h_t(x_c(t), v(t)) \end{cases} \quad (3)$$

The symbol definitions are the same as in (1). Again, we will focus only on the equation that describes the progression of the fault. It is given as:

$$x_c(t+1) = (1+\beta)x_c(t) + \omega(t) \quad (4)$$

Note that Equation (4) is a special case of the second equation in model (2). In the second equation in model (2), $x_{d,2}$=1 for a faulty condition while $x_{d,2}$=0 for a healthy one. When a fault is detected, $x_{d,2}$=1 and, therefore, they are exactly the same for a faulty condition.

3.3 Model On-line Update

Important elements in the modeling include a time-varying parameter β and noise terms ω and v. The parameter β desribes the fault growth according to the prevailing system operating conditions while the noise terms ω and v, to a certain extent, describe the confidence on the model. If a good model is developed, they can each be selected as a very small value. On the other hand, if a very rough model is used, they need to be selected each as a large value. The trade-off is that when a large noise model is used, the estimated results tend to be noisy too. A variety of techniques are available to manage effectively uncertainty [9]. In this paper, we focus on the development of the model, while noise models ω and v are determined by trial-and-error.

The parameter β depends on the loading profile that is being applied to the component of interest. Paris' Law, as shown in (5), is a relationship of fatigue crack growth under a stress intensity regime, i.e.,

$$\beta = \frac{da}{dR} = C(\Delta K)^m \tag{5}$$

Equation (5) describes the growth increment da of fault dimension per cycle dR, a is the crack length, R is the number of rotating cycles, C and m are material constants, and ΔK is the stress intensity factor. Note that for our purposes, a relationship to obtain the stress intensity factor from the loading profile and crack dimension is not available. Usually, it requires a detailed Finite Element Analysis model to reach this relationship [6]. To develop a simple, effective and affordable solution, we modify Equation (5) to arrive at the following expression:

$$\dot{L} = \frac{dL}{dt} = C_L(L)^m \tag{6}$$

stating that the rate of defect growth is related to the instantaneous crack length L under a steady operating condition. Changes in the operating conditions are reflected through parameters C_L and m, which are determined by an online adaptation routine.

When the fault mode is corrosion and/or spalling, the fault dimension is often measured by the area of the corroded or spalled surface [7]. Therefore, Equation (5) is modified as:

$$\dot{D} = \frac{dD}{dt} = C_D(D)^n \tag{7}$$

Here, C_D and n are determined by an online adaptation routine. Based on Equations (6) and (7), a defect growth model can be written as

$$L(t+\Delta t) = L(t) + \Delta t C_L(L)^m \tag{8}$$

and

$$D(t+\Delta t) = D(t) + \Delta t C_D\left(D(t)\right)^n, \tag{9}$$

respectively. Since Equations (8) and (9) have the same form, only (9) is considered in the sequel.

To achieve the goal of parameter adjustment, an adaptive prediction scheme for the bearing is given in Figure 3:

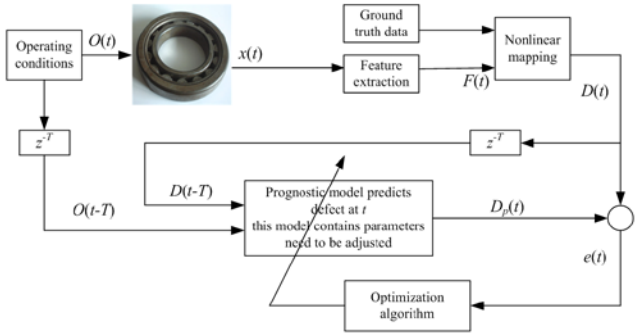

Figure 3. The on-line adaptation of prognostic model

In this model, feature $F(t)$ is extracted from the collected vibration signal $x(t)$. This information, combined with ground truth data about the defect area, is used to build a nonlinear mapping between the defect area and the feature. From this nonlinear mapping, the defect area at current time instant, $D(t)$, can be estimated. Additionally, the estimated defect area, $D(t)$, is traced back to time instant $t-T$, resulting in $D(t-T)$. This defect area $D(t-T)$ and the operating conditions serve as the input to the prognostic model to estimate the defect area at time instant t, denoted as $D_p(t)$. Then, $D(t)$ and $D_p(t)$ are compared to compute an error $e(t)$. Optimization methods can be introduced at this step to adjust the model parameters in order to minimize $e(t)$. Note that in this method, there are two preconditions: ground truth data are available to build a nonlinear mapping between features and defect area and an optimization routine is called upon to estimate the model parameters.

Note that the progression of the defect area under tightly controlled conditions could show significantly different behaviors. Therefore, the previous deterministic model must be modified to take into consideration this situation. Theoretically, the uncertainty is due to the stochastic characteristics of Paris' Law and, therefore, it is reasonable to add a random variable into the expression for the law. In practice, adding a random variable into Paris' Law is the same as adding a random variable into its parameters, i.e.,

$$\begin{aligned} D(t+\Delta t) &= D(t) + \Delta t C_D\left(D(t)\right)^n \\ C_D &= C_D + \omega_{C_D} \\ n &= n + \omega_n \end{aligned} \tag{10}$$

where C_D and n can be regarded as states associated with the model, ω_{C_D} and ω_n are zero mean random noise.

With unit step size, Equation (10) can be modified as:

$$\begin{aligned} D(t+1) &= D(t) + p_1(t) C_D\left(D(t)\right)^{p_2(t)n} \\ C_D &= C_D + \omega_{C_D} \\ n &= n + \omega_n \end{aligned} \tag{11}$$

Thus, two parameters $p_1(t)$ and $p_2(t)$ are introduced to facilitate the online parameter adaptation scheme.

A recursive least square algorithm [13] with a forgetting factor is employed to determine the parameters since it is generally fast in its convergence. The algorithm is implemented as follows:

Step 1: define a cost function as:

$$J(\theta)=\frac{1}{2}\sum_{t=1}^{T}\lambda^{T-t}\left[D(t)-D\left(\hat{\theta}(t-1)\right)\right]^{2} \quad (12)$$

where λ is the forgetting factor, which is usually given in the range of $0<\lambda\leq 1$, and $\hat{\theta}=[p_1(t)\ p_2(t)]^{T}$ is the parameter vector to be determined. In this equation, $D(t)$ and $D(\hat{\theta}(t-1))$ are ground truth data and estimate fault dimension based on model parameters, respectively.

Step 2: Calculate the derivatives of $D(\hat{\theta}(t))$ with respect to parameters θ:

$$\phi(t)=\frac{dD(\hat{\theta}(t))}{d\theta} \quad (13)$$

Step 3: The parameter update is given by:

$$\hat{\theta}(t)=\theta(t-1)+P(t)\phi(t)\left[D(t)-D\left(\theta(t-1)\right)\right] \quad (14)$$

and $P(t)$ is updated as

$$P(t)=\frac{P(t-1)}{\lambda}\left[1-\frac{\phi(t)\phi^{T}(t)P(t-1)}{\lambda+\phi^{T}(t)P(t-1)\phi(t)}\right] \quad (15)$$

The recursive least square method with a forgetting factor actually applies an exponential weighting term to the past data. In the cost function (12), the influence of past data reduces gradually as new data become available and the algorithm can be easily implemented on-line.

To implement the algorithm, a set of initial parameters must be given. Parameter $\theta(0)$ is specified from our prior knowledge of the system while $P(0)$ is written as a large number times an identity matrix.

Note that the previous parameter adaptation is realized by a recursive least square method. Other methods, such as an extended Kalman filter [13], a neural network [12], etc., can be used as well.

3.4 Consideration of Operating Conditions

In the previous model, the operating conditions, such as ambient temperature, humidity, load, grease quality, etc., are not taken into consideration. The influence of the operating conditions is reflected by the on-line parameter tuning. If the operating conditions can be compensated, a precise fault progression model can be derived.

Let us consider the corrosion fault mode. It is known that humidity is a leading contributing factor of corrosion. Therefore, the environmental humidity should be incorporated in the corrosion progression model. Suppose that the nominal humidity (could be normal room humidity) is denoted by H_n. The environmental humidity is measured as H_c. The normalized humidity condition then can be described as a humidity index h_i given by $h_i=H_c/H_n$. Clearly, large humidity values result in larger h_i, while small humidity values result in smaller h_i.

If we know that the humidity influences linearly the corrosion propagation, the previously modified Paris' law (7) can be further re-written as (16) to include the humidity factor.

$$\dot{D}=\frac{dD}{dt}=h_i C_D(D)^{n} \quad (16)$$

If, however, we know that humidity influences exponentially the corrosion progression, (7) can be re-written as:

$$\dot{D}=\frac{dD}{dt}=C_D(D)^{n\cdot h_i} \quad (17)$$

Accordingly, the discrete time model should be modified as:

$$D(t+1)=D(t)+h_i C_D\left(D(t)\right)^{n}, \quad (18)$$

and

$$D(t+1)=D(t)+C_D\left(D(t)\right)^{n\cdot h_i}, \quad (19)$$

respectively.

When other operating conditions are considered, such as nominal temperature T_n (normal operating temperature under healthy conditions), nominal load L_n (rated load), and grease quality G_n, (will be discussed later), appropriate terms in the model can be defined as well. Then, a temperature index t_i, a load index l_i, and a grease quality index g_i can be determined. Multiple ways are available to combine them into a single model.

Suppose these factors influence linearly the fault progression, a possible alternative is to write the model as either

$$D(t+1)=D(t)+h_i t_i l_i g_i C_D\left(D(t)\right)^{n}, \quad (20)$$

or

$$D(t+1)=D(t)+\left(w_h h_i+w_t t_i+w_l l_i+w_g g_i\right)C_D\left(D(t)\right)^{n}, \quad (21)$$

where w_h, w_t, w_l, w_g are weighting factors and $w_h+w_t+w_l+w_g=1$.

The model can be written as either

$$D(t+1)=D(t)+C_D\left(D(t)\right)^{nh_i t_i l_i g_i}, \quad (22)$$

or

$$D(t+1)=D(t)+C_D\left(D(t)\right)^{n\left(w_h h_i+w_t t_i+w_l l_i+w_g g_i\right)}, \quad (23)$$

when the factors are exhibiting an exponential dependence.

It is possible that the influence of some factors is exhibiting a linear behavior, while that of others is exponential. In this case, suppose the linear factors are $f_{l,1}$ and $f_{l,2}$ and the exponential ones are $f_{e,1}$ and $f_{e,2}$. Then, the model can be written as

$$D(t+1)=D(t)+f_{l,1}f_{l,2}C_D\left(D(t)\right)^{nf_{e,1}f_{e,2}}, \quad (24)$$

or

$$D(t+1)=D(t)+\left(w_{l,1}f_{l,1}+w_{l,2}f_{l,2}\right)C_D\left(D(t)\right)^{n\left(w_{e,1}f_{e,1}+w_{e,2}f_{e,2}\right)}, \quad (25)$$

where $w_{l,1}$, $w_{l,2}$, $w_{e,1}$, $w_{e,2}$ are weighting factors and $w_{l,1}+w_{l,2}=1$ and $w_{e,1}+w_{e,2}=1$.

Grease Quality

Unlike temperature, humidity, and load, whose nominal values can be set as temperature under healthy condition, room humidity and rated load, grease quality is difficult to quantify. It is known that the quality of grease can vary significantly and changes in grease must be accounted for in the model. It is further aggravated due to material peeling and debris accumulation – the byproducts of spalling and corrosion. Thus, the grease quality and its dependence on corrosion/spalling may be quantified as shown in Figure 4

(a). The index of grease quality g_i can be illustrated as in Figure 4 (b).

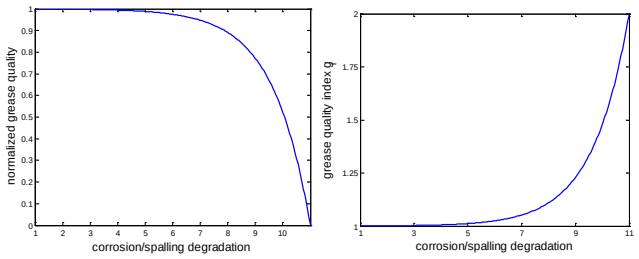

Figure 4 Illustration of grease quality degradation (a) Grease quality degradation vs. corrosion/spalling; (b) grease quality index vs. corrosion/spalling

The initial grease quality index g_i can be normalized to a value of 1. Then, the grease quality decreases exponentially while the grease quality index increases exponentially as corrosion/spalling develops. This increased grease quality index, will accelerate the fault growth as reflected in the model.

3.5 An Extended Model

From the above descriptions, it is clear that there are usually more than one fault modes present in a failing bearing. Although a corrosion model can describe the progression of corrosion, a spalling model can describe the progression of spalling, and a cracking model can describe the progression of cracking, an extended model to describe a "combined" or "fused" health condition is desired. Since these different fault modes work together in a coupled fashion, this "combined" or "fused" health condition could provide a more precise and accurate evaluation of the health state of the bearing.

To this end, a generalized degradation (fault) variable is defined as g. Then, the progression of this generalized degradation variable can be described as

$$\frac{dg}{dR} = f\left[b_1 w_1(corrosion) + b_2 w_2(spalling) + b_3 w_3(cracking)\right] \quad (26)$$

where w_1, w_2, w_3 ($w_1+w_2+w_3=1$) are weighting factors for different fault modes, b_1, b_2, b_3 are time varying parameters indicating that corrosion, spalling, and cracking are detected and their respective prognostic module is activated.

To use this model in diagnosis and prognosis, we must consider the fault features and the available ground truth data. In Equation (26), the different fault modes are weighted and summed. The feature vectors, for different fault modes, can be fused via intelligent methods, such as genetic algorithms, genetic programming, Dempster-Shafer theory, neural networks, fuzzy logic, Kalman filtering, etc., instead of simply adding their contributions to the degradation formula.

4 EXPERIMENTAL RESULTS

4.1 Multi-Fault Modes in-situ Data

Three sets of vibration data acquired from bearings are used to demonstrate the approach. Each data set corresponds to a different fault mode. The three fault modes represented are grease breakdown of a helicopter bearing, spalling of a test bearing, and an unknown fault from a bearing on a helicopter. From each data set, an adequate feature is extracted to indicate the progression of its corresponding fault.

In practice, the data are often available at very limited service time instants. Such is the case in the example data used. To implement the fault detection and failure prognosis algorithms, more data points must be generated. A simple way is to interpolate between the available data.

The original and interpolated feature values are then normalized in the range of [0 1] to facilitate the initialization of parameters. Additional benefits of the normalization operation are to minimize the effect of ill-conditioning and to provide more stable inputs to the fault diagnosis and failure prognosis modules. To simulate the real cases, a white noise term is added to the normalized feature data.

4.1.1 Fault 1: Grease breakdown

The data under study consists of a time series of accelerometer readings from a series of tests on bearings from the oil cooling subsystem of H-60 helicopters. These experiments were conducted by industry and the U.S. Army to test bearing behaviors with grease degradation. The data of focus consist of sampled vibration files, in a time sequence, recording the vibration signal of bearing with degraded lubrication. The sampling rate is 10 KHz.

The feature extracted from this data set is the Root Mean Square (RMS) value of the vibration signal in a frequency band of 2 to 4 KHz. The data set contains 24 files at different service times. The 24 points are interpolated in time to simulate progressive degradation.

4.1.2 Fault 2: Spalling from bearing experiments

In this case, vibration data were acquired from a bearing with a naturally occurring spall at a sampling frequency of 204.8 kHz. The vibration data are acquired at 4 different times from a test that ran the bearing for just over 16 hours. Each vibration segment contains 8 seconds of vibration. The bearing was disassembled and reassembled 3 times throughout the test to measure the fault dimension.

In this experiment, bearing failure begins with a spall developing on the surface of a raceway. As the spall grows, it excites increasingly a specific frequency associated with this type of defect. The amplitude and time duration of the frequency excitation are indicative of defect severity.

Features are extracted from an enveloped / modulated segment of the vibration signal in the frequency domain. The feature values are the sum of weighted frequency components related to harmonics of the frequency of interest. As mentioned above, only 4 data segments, indicative of the health condition, are available at different service hours. From this data set, four feature values are extracted and then interpolated to a rate of one feature value per minute.

4.1.3 Fault 3: Unknown fault from bearing on-board helicopter

These data were generated from vibration signals acquired from an accelerometer on-board a helicopter that monitored

the condition of a bearing for about two-and-a-half years. The data set contains 420 data files. Each data file corresponds to one of three different helicopter operating regimes: on-ground, hover or level flight. The exact fault mode remains unknown, but the bearing was removed after its condition was deemed to be suspect per the corresponding monitoring activities.

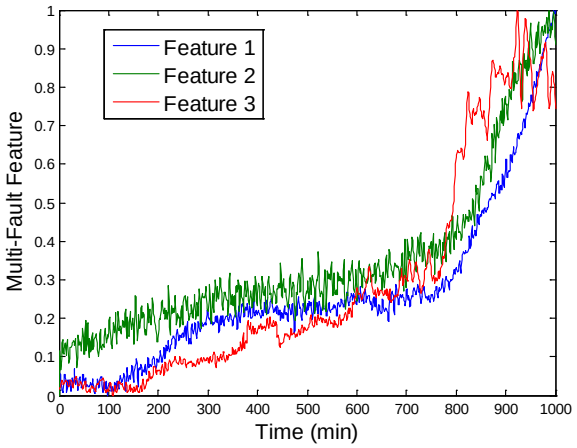

Figure 5 Interpolated features for multi-fault modes

4.2 Experimental Results

The feature values of the three fault modes described are shown in Figure 5. In this figure, the feature vectors have been interpolated on the time axis such that they have comparable time scales so as to allow for the demonstration of the diagnostic and prognostic algorithms. The feature vectors are also normalized and noise is added.

We now use the three feature value series to simulate a system with 3 fault modes occurring simultaneously. For convenience, we further assume that the end of the useful life of the system is defined at the end of our data series.

To implement the diagnostic and prognostic algorithms in our "combined" system, the fault progression model must be initialized. Parameter C and n for each fault mode are given from our prior knowledge of the system while $\theta(0)$ is given as 1 and $P(0)$ is specified as a large number times an identity matrix. To implement the algorithm, the initial parameters are set as: $[C_1, n_1] = [5,2, 1]$, $[C_2, n_2] =[2.0, 1]$, $[C_3, n_3] =[3.8, 1]$ and $\theta(0)=[1, 1]$, and $P(0)=[100\ 0; 0\ 100]$ for all three fault modes.

For fault detection, 500 particles are used, while for failure prognosis and in order to reduce the computational burden, 30 particles are used.

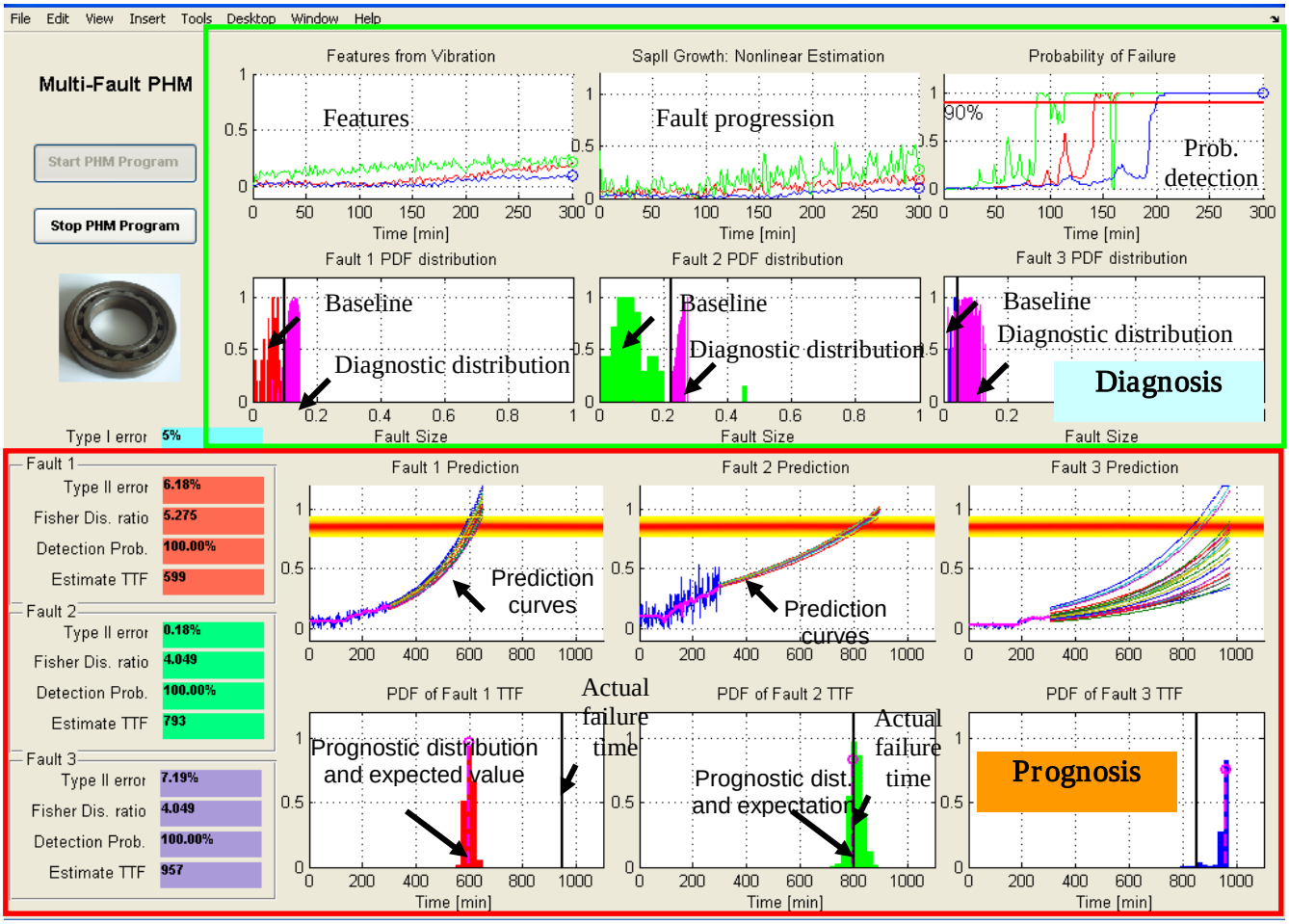

Figure 6. Diagnostic and prognostic program

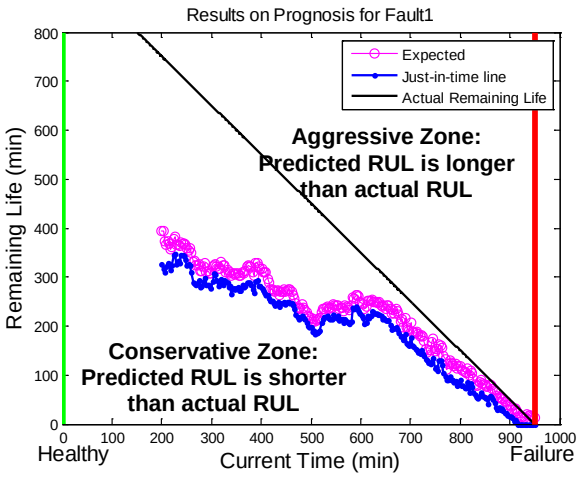

Figure 7. Estimated and actual RUL for fault 1

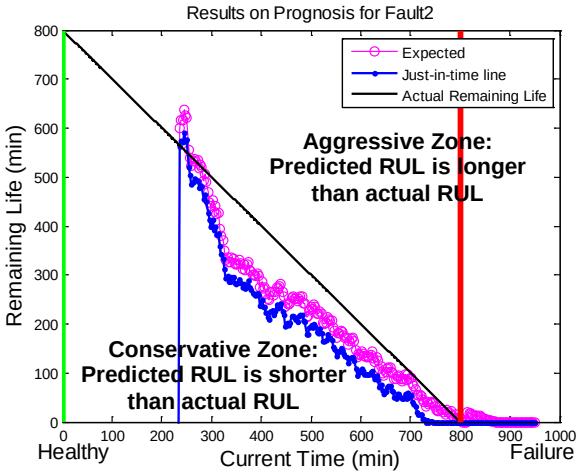

Figure 8. Estimated and actual RUL for fault 2

The experimental results on the interpolated data are shown in Figure 6 to Figure 9. Figure 6 shows the implementation of the program. In the diagnosis box, which is on the top of Figure 6, the features, fault progression curves, and probability of detection are shown in three sub-figures. The three subfigures below them are real-time distributions of the three faults, respectively. In the three figures, baseline distributions and the value of $z_{1-\alpha,\mu,\sigma}$ are shown and compared with the real-time distribution. Here, $z_{1-\alpha,\mu,\sigma}$ is the threshold value for a normal distribution with mean value μ, standard deviation σ and α 5%. Thus, it is an indicator for detection when the sum of the weights of all particles is larger than this threshold value.

In the prognosis box, the top three subfigures shows the prediction curves for three different fault modes with their respective hazard zones. The bottom three subfigures show the comparison of prognostic distribution and its expected value to actual failure time, which is available when the fault dimension reaches a predefined value. The leftmost frames show the indicators of fault detection and failure prognosis, which include type II error, the Fisher's Discriminant Ratio, probability of detection, and estimate of time to failure (TTF) [4].

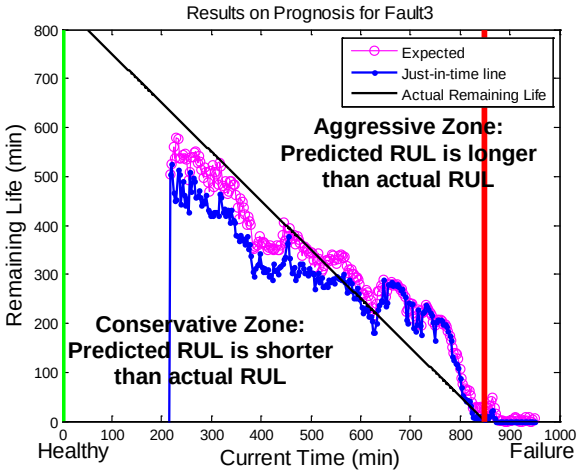

Figure 9. Estimated and actual RUL for fault 3

At the beginning, only the diagnosis routine is activated while the prognosis routine remains inactive. When a fault mode is detected, the prognostic part is activated. The detection and prognosis of three fault modes run in parallel. The real-time diagnostic distribution is used as the initial condition for failure prognosis. When a new measurement becomes available, the activated prognostic algorithm calculates the long-term prediction for each particle and the time instances that these particles reach the predefined hazard zone to form a real-time distribution of RUL.

The expected value of the estimated RUL pdf from the time of detection to the end of data, the lower bound of the 95% confidence interval (Just-in-time line [14]), and the actual remaining useful life for three fault modes are shown from Figure 7 to Figure 9. Note that the area below the actual remaining useful life is the conservative estimation zone. In practice, an estimate of the RUL in the conservative zone is desired because this means that the estimated RUL is shorter than the actual RUL. This will lead to early maintenance but will not place the bearing's safety at risk.

5 CONCLUSION

The emerging Condition-Based Maintenance and Prognostics and Health Management technologies promise to deliver substantial benefits in terms of improved reliability, availability, safety and maintainability for complex engineered systems. The development and implementation of these technologies though is presenting major challenges to the system designer. Model-based fault diagnosis and failure prognosis has demonstrated significant advantages over

other methods, with the key to successful implementation being the availability of a model that represents faithfully the fault mode progression. Moreover, critical system components may exhibit various fault mode stages during their operating life. The accurate representation of the fault stages in terms of a suitable modeling framework is essential in the development of reliable and robust diagnostic and prognostic algorithms. This paper is attempting to motivate the need for multi-fault modeling approaches and to illustrate their efficacy with examples from the rotating machinery world.

ACKNOWLEDGMENT

This work was supported by the Army Research Laboratory to Impact Technologies, LLC. We acknowledge their support and collaboration in the conduct of this research effort.

REFERENCES

[1] P. McFadden and J. Smith, "Model for the vibration produced by a single point defect in a rolling element bearing" , Journal of Sound and Vibration, vol. 96, pp. 69-82, 1984.

[2] I. Howard, "A review of rolling element bearing vibration: detection, diagnosis and prognosis", DSTO-RR-0013, Airframes and Engines Division, 1994.

[3] B. Li, M.-Y. Chow, Y. Tipsuwan and J. Hung, "Neural networks based motor rolling bearing fault diagnosis", IEEE Transactions on Industrial Electronics, vol. 47, no. 5, pp 1060-1069, Oct. 2000

[4] G. Vachtsevanos, F. Lewis, M. Roemer, A. Hess and B. Wu, "Intelligent fault diagnosis and prognosis for engineering systems", Wiley, 2006

[5] M. Orchard and G. Vachtsevanos, "A particle filtering based framework for real-time fault diagnosis and failure prognosis in a turbine engine", 15th Mediterranean Conference on Control and Automation, Athens, Greece, July 2007.

[6] R. Patrick, A Model Based Framework for Fault Diagnosis and Prognosis of Dynamical Systems with an Application to Helicopter Transmissions, Ph.D. Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, July, 2007.

[7] Y. Li, S. Billington, C. Zhang, T. Kurfess, S. Danyluck, S. Liang, "Adaptive prognosis for rolling element bearing condition" Mechanical Systems and Signal Processing, 13(1), pp.103-113, 1999.

[8] M. Orchard, G. Kacprzynski, K. Goebel, B. Saha, G. Vachtsevanos, Advances in Uncertainty Representation and Management for Particle Filtering Applied to Prognostics, International Conference on Prognostics and Health Management, Oct. 2008, Denver CO, USA.

[9] B. Randall, J. Antoni, S. Chobsaard, The Relationship Between Spectral Correlation and Envelope Analysis in the Diagnostics of Bearing Faults and other Cyclostationary Machine Signals, Mechanical System and Signal Processing, 15(5), 2001, pp 945-962.

[10] P. Tse, Y. Peng, R. Yam, Wavelet Analysis and Envelope Detection for Rolling Element Bearing Fault Diagnosis - Their Effectiveness and Flexibilities, J. Vibration and Acoustics, 123(3), 2001, pp 303-310.

[11] D. Ho, B. Randall, Optimisation of Bearing Diagnostic Techniques Using Simulated and Actual Bearing Fault Signals, Mechanical System and Signal Processing, 14(5), 2000, pp 765-788.

[12] P. Goode, M.-Y. Chow, Using a neural/fuzzy system to extract heuristic knowledge of incipient faults in induction motors. Part I-Methodology, IEEE Transactions on Industrial Electronics, 42(2), 1995, pp 131-138.

[13] P. Zarchan, H. Musoff, Fundamentals of Kalman Filtering A Practical Approach, Progress in astronautics and aeronautics, v. 208. Reston, Va: American Institute of Aeronautics and Astronautics, 2005

[14] S. Engel, B. Gilmartin, K. Bongort, A. Hess, Prognostics, the Real Issues Involved With Predicting Life Remaining". Proceedings of the IEEE Aerospace Conference, Big Sky, Montana, March 18-25, 2000.

[15] C., Zhang, T. Kurfess, S. Danyluk, S. Liang, Dynamic modeling of vibration signals for bearing condition monitoring, the 2nd International Workshop on Structural Health Monitoring, Stanford, 1999, 926-935.

[16] Y. Choi, C. Liu, Rolling contact fatigue life of finish hard machined surfaces Part I. Model develop, Wear, 261, 2006, 485-491.

[17] M. Davies, Y. Chou, C. Evans, on chip morphology, tool wear and cutting mechanics in finish hard turning, Ann. CIRP 45(1), 1996, 77-82.

[18] E. Ioannides and T. Harris, A new fatigue life model for rolling bearing, Trans. ASME, J. Tribology, 1985, 107, 367-278.

[19] J. Qiu, C. Zhang, B. Seth, and S. Liang, Damage mechanics approach for bearing lifetime prognostics, Mechanical Systems and Signal Processing, 16(5), 2002, 817-829.

[20] D. He and E. Bechhoefer, Bearing Prognostics using HUMS condition indictors, American Helicopter Society 64th Annual Forum, Montreal, Canada, 2008.

[21] M.N. Kotzalas, T.A. Harris, Fatigue failure progression in ball bearings, J. Tribology 123 (2001) 238–242.

[22] C.J. Li, H. Shin, Tracking bearing spall severity through inverse modeling, Proceedings of the ASME International Mechanical Engineering Congress, Anaheim, CA, USA, 2004, pp. 1–6.

Bin Zhang received the B.E. and M.S.E. degrees in mechanical engineering from Nanjing University of Science and Technology, Nanjing, China, in 1993 and 1999, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2007. He was been a Postdoctoral Researcher in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA before he joined Impact Technologies, LLC, Rochester, NY. He is the author and coauthor of more than 70 technical papers. His current research interests include fault diagnosis and failure prognosis, systems and control, digital signal processing, learning control, intelligent systems and their applications to robotics, power electronics, and various mechanical systems.

Chris Sconyers is a PHD student at the Georgia Institute of Technology. He received his B. S. in Electrical Engineering and his B. S. in Computer Science from Texas A&M University in 2004, and his M. S. in Electrical and Computer Engineering from the Georgia Institute of Technology in 2006. His research interests include component-level diagnostics and prognostics, and simulation, perception, tracking, and control for unmanned aerial vehicles.

Romano Patrick received the degrees in electrical engineering from the University of Texas, Arlington, and the University of Panamericana, Guadalajara, Mexico, and the M.B.A. degree and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, in 2007. He is currently a Project Manager with Impact Technologies, LLC, Rochester, NY. His current research interests include interdisciplinary integration of hardware, software, and techniques to support cost-effective design and implementation of engineering systems. He was involved in electronic, mechanical, and software design, state-of-the-art process automation, and novel machine health monitoring for a variety of

industrial and government sponsors. He was a Graduate Professor and a Program Coordinator at the University of Panamericana.

George Vachtsevanos received the B.E.E. degree in electrical engineering from the City College of New York, New York, NY, in 1962, the M.E.E. degree in electrical engineering from New York University, New York, in 1963, and the Ph.D. degree in electrical engineering from the City University of New York, New York, in 1970. He is currently a Professor Emeritus of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, where he directs the Intelligent Control Systems Laboratory. His work is funded by government agencies and industry. He is the author or coauthor of more than 240 technical papers.