Fault Progression Modeling: An Application to Bearing Diagnosis and Prognosis

Bin Zhang, Chris Sconyers, Marcos Orchard, Romano Patrick, George Vachtsevanos

Abstract—The successful implementation of fault diagnosis and failure prognosis algorithms to safety critical systems requires the definitions and applications of mathematically rigorous modules. These modules, including data pre-processing, feature extraction, diagnostic and prognostic algorithms, performance metrics definition, and a fault progression model, form an integrated architecture for system health monitoring and management. In these modules, the fault progression model is critical to detection of incipient failures as early as possible with predefined specifications and prediction of the system’s remaining useful life accurately and precisely. This paper considers an oil cooler bearing of a helicopter and proposes a methodology for fault detection and failure prognosis, in which data pre-processing, feature extraction and fault progression modeling are discussed in detail. Experimental results are presented to verify the proposed methodology and fault progression model.

Index Terms—Fault detection, Failure prognosis, Fault progression modeling, Feature extraction, Data processing

I. INTRODUCTION

ROLLING element bearings are critical components in rotating machinery to permit constrained relative rotation or linear motion between two parts. Bearings are widely used and studied extensively over the past years [1-3]. Due to high loading and severe operating conditions and environments, defects are often developed on the bearing components after a period of service time. Bearing faults such as corrosion, spalling and pitting on the raceway surfaces will increase friction causing overheating and may lead to failure of the bearing by expansion and freezing of the rotating machinery. The published literature elaborates about fault detection of different fault modes including spalls, pits, misalignment and waviness on rolling surfaces [4, 12-14].

The capability to detect, isolate and identify incipient failures of bearings is critical to determining appropriate mitigation actions to maintain system safety and reliability [5]. Moreover, such diagnostic capabilities must be accompanied with the ability to predict accurately and precisely the Remaining Useful Life (RUL) of failing bearings in order for the system operators to take corrective action in the event of safety critical failures.

Recent advances in sensing, monitoring and fault diagnosis have resulted in the development and application of health and usage monitoring systems that acquire and analyze pertinent data on-line. Although fault detection, isolation and identification technologies have been demonstrated to a satisfactory degree, prognostics, i.e. the ability to predict accurately the RUL of failing components, presents major challenges to the Prognostics and Health Management (PHM) designer due primarily to the difficulty to model the fault progression and to account for the inherent uncertainty associated with the task of long-term prediction [11].

In this paper, we introduce an integrated methodology and tools for the fault diagnosis and failure prognosis of bearings. The proposed architecture is composed of mathematically rigorous modules that are applicable not only to bearings but also to other critical aircraft components. The emphasis is to integrate efficiently and effectively modules including signal pre-processing, feature extraction and fusion, fault progression modeling and parameter adaptation, diagnostic and prognostic algorithms. In the proposed methodology, vibration signals, collected at different phases of a bearing, with degradation of their health status are processed to remove noise and reduce signal variability. Features are then extracted. These features or characteristic signatures of the fault signal reduce significantly the data dimensionality without sacrificing the signal’s information content. They form the foundation for “good” fault detection and failure prognosis algorithms.

An accurate fault progression model for the components under consideration describing the fault growth or system state degradation is important for accurate and precise detection and prognosis. For fault detection, the particle filtering based [6] diagnostic algorithm generates an on-line estimation of the feature and fault dimension distributions when new measurements become available. Statistical analysis then can be applied by comparing the deviation between the real-time distribution from the algorithm and the baseline data distribution to arrive at the probability of fault condition and false alarm rate. As soon as a fault is detected, a failure prognostic algorithm is activated to predict the RUL of the bearing or the time that the fault dimension reaches a certain hazard level.

In the following sections, the system and bearing under consideration are presented briefly in Section II. The data, data pre-processing and feature extraction tools are discussed in Section III. Modeling of the fault progression for diagnosis and prognosis is addressed in Section IV. Experimental results are presented in Section V, which is followed by conclusions.
II. SYSTEM UNDER CONSIDERATION

The oil cooler is a core component of the H-60 drivetrain assembly. Its primary function is to cool the helicopter transmission lubricant while transmitting power to the tail rotor drive shaft through the oil cooler shaft. It sits in the downdraft of the main rotor wash, channeling the flow of air through the heat exchanger for efficient cooling. In physical terms, the tail rotor drivetrain consists of a drive shaft that transfers torque from the main transmission to the oil cooler drive shaft, four drive shaft interconnected sections that transfer torque from the oil cooler drive shaft to the intermediate gear box, and another drive shaft that transfers torque from the intermediate gear box to the tail rotor gear box. This arrangement is shown in Figure 1.

The components of the oil cooler are centered around a splined shaft supported at the front by two shielded cartridge bearings and in the rear by a viscous damper bearing. The oil cooler fan assembly consists of a multi-bladed rotor housed inside a concentric stator. Shielded cartridge type bearings are used where space for lubrication is limited or when operating conditions demand a larger grease supply inside the bearings. Fault analysis of the oil cooler is critical since a failure to effectively regulate drive train temperature may eventually lead to expansion and freezing of the drive shaft and may result in catastrophic power loss. The bearing supporting the cooler fan is the component most prone to failure and, therefore, of interest in this paper. The dominant fault modes include grease breakdown, debris ingestion, etc. When the vehicle is operating in a high moisture environment, corrosion is also a significant contributing factor to failure because it will lead to spalling and cracking.

III. DATA PRE-PROCESSING AND FEATURE EXTRACTION

A. Vibration data

Vibration data were acquired from a bearing with a naturally occurring spall at a sampling frequency of 204,800 Hz. The vibration data are measured at different bearing service times. Each segment contains data of 8 seconds period. During the operation, the bearing is disassembled after a number of certain services hours to investigate the health condition of the bearing. The data and health conditions are listed in Table 1.

Table 1 Vibration Data and Fault Dimension

<table>
<thead>
<tr>
<th>Bearing service time</th>
<th>Data segment No.</th>
<th>Data length</th>
<th>Spall area (mm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>1.5 hrs</td>
<td>DN1</td>
<td>8 s</td>
<td>NA</td>
</tr>
<tr>
<td>3.5 hrs</td>
<td>DN2</td>
<td>8 s</td>
<td>NA</td>
</tr>
<tr>
<td>8 hrs</td>
<td>NA</td>
<td>NA</td>
<td>0.3x0.2</td>
</tr>
<tr>
<td>12 hrs</td>
<td>NA</td>
<td>NA</td>
<td>0.4x0.2</td>
</tr>
<tr>
<td>13 hrs</td>
<td>DN3</td>
<td>8 s</td>
<td>NA</td>
</tr>
<tr>
<td>15.5 hrs</td>
<td>DN4</td>
<td>8 s</td>
<td>NA</td>
</tr>
<tr>
<td>16 hrs</td>
<td>NA</td>
<td>NA</td>
<td>0.5x0.2</td>
</tr>
</tbody>
</table>

B. Data pre-processing

In practice, the collected vibration signals are often corrupted by a wide-band noise, which will degrade the quality of features and will, eventually, increase the false alarm rate while decreasing the accuracy and precision of failure prognosis. Improvements in terms of increased signal-to-noise ratio and reduced signal variability contribute to higher detection confidence, smaller false alarms rates, and accurate and precise prognosis.

The vibration measurement can be described as:

\[ s(t) = v_s(t) + n(t) \]  

where \( s(t) \) is the measured signal, \( v_s(t) \) is the true vibration signal, which can be considered as the sum of narrow band signals spaced by a fault characteristic frequency, and \( n(t) \) is broad band noise. It is known that when a fault, for example, a spall on the raceway surface, is initiated in a bearing, an impulse of vibration is generated when the rolling elements pass through the spall. This impulse is normally repetitive because the rolling elements contact the defect on the bearing surface in a periodic manner. The impulse generated by the defect is much shorter than the bearing rotating period. Therefore, the energy of the impulse will be distributed across a very wide frequency range which will excite various resonances of the bearing [2]. The frequency of the impulses is usually regarded as the bearing characteristic defect frequency and is critical for fault detection purposes.

Since adaptive line enhancement techniques [8, 10] are widely used to separate narrow band signals from broad band noise, these algorithms are applied to bearing vibration data in the experiment. The adaptive line enhancement scheme is shown in Figure 2. In the scheme, the delayed measurement \( s(t-\Delta) \) is used as the reference signal and is fed into an adaptive filter. The output of the filter is compared with the measurement, \( s(t) \), to derive the error signal \( e(t) = s(t) - F(s(t-\Delta)) \), where \( F(.) \) is the adaptive filtering process. The error signal is then used to tune the parameters of the filter via a least mean square method [16].

As shown in Figure 2, in this adaptive line enhancement scheme, the correlation of the broad band noise \( n(t) \) in the delayed reference signal \( s(t-\Delta) \) and the measurement \( s(t) \) is...
removed by the delay $\Delta$. However, the correlation of the true vibration signal $v_b(t)$ in these two signals is still high. By minimizing the error, the filter tends to keep the correlated narrow band true vibration signal while it cancels the un-correlated broad band noise. The adaptive filter will eventually converge to a band-pass filter which has a pass band centered by the frequencies of the narrow band true vibration signal.

![Figure 2. Adaptive line enhancement scheme](image)

After the broad band noise is removed, the signals may be further processed to highlight the fault signature. Envelope analysis, by demodulating the vibration signals at the resonances that the impulse is exciting, not only detects the presence of a defect, but also localizes the defective component [2, 12-14]. It provides a mechanism for extracting the periodic excitation of the resonance from the vibration signals. The frequency of the extracted signal is the frequency of the impulse, i.e., the characteristic bearing defect frequency.

### C. Feature extraction and interpolation

Features are extracted from the envelope signal in the frequency domain and the scheme is illustrated in Figure 3. In this example, the frequency band around a fault characteristic frequency and its harmonics are weighted and summed. First, a weighting window in the frequency domain is defined, which has the same form for all harmonics. The window is usually selected as a Gaussian or Hamming shape. The length of the window depends on how many frequency components will be included in the feature calculation. The feature of interest is the sum of all frequency components in the weighted frequency bands.

![Figure 3. Feature extraction.](image)

As mentioned in Table 1, only 4 data segments indicative of the health condition are available at different service hours. From this data set, only four features are extracted. To implement the fault detection and failure prognosis algorithms, more data points must be generated. In the absence of additional experimental data, the simplest way is to interpolate the available data. Note that the data given in Table 1 are at different hours. Therefore, the feature and ground truth fault dimension can be interpolated at the time unit of a minute.

![Figure 4. Interpolation of fault dimension in unit of minute](image)

![Figure 5. Interpolation of feature values (noise is added)](image)

![Figure 6. Nonlinear mapping of fault progression according to feature vector](image)

Figures 4 and 5 show the interpolated data of the ground truth fault dimension and feature value, respectively. Note that the feature is normalized in order to minimize the effect of ill-conditioning and to provide more stable inputs to the fault diagnosis and failure prognosis module. The z-score method is employed to achieve this task:
\[ X = \frac{(Z - \mu)}{\sigma} \]  
where \( X \) is the normalized feature, \( Z \) its original feature and \( \mu, \sigma \) the feature’s mean and standard deviation, respectively. A white noise is added to the interpolated feature data. Then, from these two figures, a nonlinear mapping of the fault progression as a function of the feature values can be estimated and is shown in Figure 6. These interpolated data are used to verify the proposed methodology.

IV. DIAGNOSIS AND PROGNOSIS MODELS

The fault progression is often nonlinear and, consequently, the model is 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 [9]. An essential element of this module is a nonlinear state model describing the progression of the fault.

A. 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. A model for diagnosis is given as:

\[
\begin{align*}
\dot{x}_d(t + 1) &= f_b\left(x_d(t) + n(t) \right) \\
X(t + 1) &= f_l(t,u,x_d(t),\chi(t),\omega(t)) \\
\text{Features}(t) &= h_x(x_d(t),\chi(t),v(t))
\end{align*}
\]  

where \( f_b, f_l, h_x \) 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, \( a(t) \) and \( v(t) \) are non-Gaussian process and feature noise signals, respectively, \( n(t) \) is i.i.d. uniform white noise, \( t \) is a time stamp and \( u \) is the input, including the loading profile, operating speed, temperature, humidity, etc. Note that the last equation in (4) builds a connection between the system state and the feature vector.

Suppose that in diagnosis, only two states, normal and faulty, are considered. Moreover, assume the nonlinear mapping \( h_x(.) \) between the feature and fault state is one-to-one, then, for an actual system, model (4) can be re-written to facilitate the implementation of a fault detection routine as:

\[
\begin{align*}
\dot{x}_d,1(t + 1) &= f_b\left(x_d,1(t) + n(t) \right) \\
\dot{x}_d,2(t + 1) &= f_b\left(x_d,2(t) + n(t) \right) \\
\chi(t + 1) &= \left[(1 + \beta(t,u))x_d(t) \right] \cdot x_d,2(t) + \omega(t) \\
y(t) &= \chi(t) + v(t) \\
f_b(x) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{if} \quad \|x - [1 & 0]\| \leq \|x - [0 & 1]\| \\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{else}
\end{align*}
\]  

in this model, \( x_{d,1} \) and \( x_{d,2} \) are Boolean states that indicate normal and faulty conditions, respectively, and \( \beta \) is a time-varying parameter that describes the progression of the fault dimension under a fatigue stress.

B. Failure Prognosis

Prognosis is activated when a fault is detected. For the same fault mode in detection and prognosis, the progression of the fault follows the same physical law. Therefore, in the prognosis model, the Boolean state can be removed from model (4) and a model for prognosis is derived as follows:

\[
\begin{align*}
\chi(t + 1) &= f_p(t,u,\chi(t),\omega(t)) \\
\text{Features}(t) &= h_x(x_d(t),v(t))
\end{align*}
\]  

The definitions of symbols are the same as in (4). Again, we will only focus on the equation that mainly describing the progression of the fault. It is given as:

\[ \chi(t + 1) = (1 + \beta(t,u))x(t) + \omega(t) \]  

Note that Equation (7) is a special case of the second equation in model (5). In the second equation in model (5), \( x_{d,2}=1 \) for a faulty condition while \( x_{d,2}=0 \) for a healthy condition. When a fault is detected, \( x_{d,2}=1 \) and, therefore, they are exactly the same for a faulty condition.

C. Model of fault progression

Important elements in the modeling include a time-varying parameter \( \beta \) and noises \( \omega \) and \( v \). The parameter \( \beta \) describes the fault growth according to system operating conditions while noises \( \omega \) and \( v \) to a certain extent, describe the confidence on the model. If a good model is developed, they can be selected as a very small value. On the other hand, if a rough model is used, they need to be selected as large values. The trade-off is that the estimated results tend to be noisy when large noises are used. In this paper, we focus on the development of the model \( \beta \). Noises \( \omega \) and \( v \) are determined by trial-and-error. On-going research on the representation and management of \( \omega \) and \( v \) can be found in [11].

The parameter \( \beta \) depends on the loading profile applying to the bearing. Paris’ Law, as shown in (8), is a relationship of fatigue crack growth under a stress intensity regime, i.e.

\[ \beta = \frac{d\alpha}{dR} = C(\Delta K)^n \]  

Equation (8) describes the growth increment \( da \) of the fault dimension per cycle \( dR \), \( a \) is the fault dimension, \( R \) is the number of rotating cycles, \( C \) and \( n \) are material constants, and \( \Delta K \) is the stress intensity factor. Note that for our purposes, a relationship representing the stress intensity factor as a function of the loading profile and crack size is not available. Usually, it requires a detailed Finite Element Analysis model to determine this relationship [7]. For a bearing, spalling is a common fault mode and it is measured by the area of spall, as shown in Table 1. To develop a simple, effective and affordable solution, we modify Equation (8) to:

\[ \dot{i} = a(D)^n \]  

that is the rate of defect growth is related to the instantaneous fault dimension \( D \), which is the area of spalling, under a
steady operating condition. When the operating condition is changed, the change is reflected through the parameters \( C_D \) and \( n \), which are determined by an online adaptation routine. Based on Equations (9), a defect growth model can be written in the discrete-time form as

\[
D(t+1) = D(t) + C_D \left( D(t) \right)^n.
\]

(10)

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 Paris’ Law. In practice, adding a random variable into Paris’ Law is the same as adding a random variable into its parameters and we arrive at:

\[
D(t+1) = D(t) + p_1(t) C_D \left( D(t) \right)^{p_2(t)n} + \omega_c + \omega_n
\]

(11)

where \( C_D \) and \( n \) can be regarded as states associated with the model, and \( \omega_c \) and \( \omega_n \) are zero mean random noise. Two parameters \( p_1(t) \) and \( p_2(t) \) are introduced to facilitate the online parameter adaptation scheme.

D. Model parameter online adaptation

To determine the parameters, a recursive least square algorithm [16] with a forgetting factor is employed since it is generally fast in its convergence. The algorithm is implemented as follows:

Step 1: define a cost function as:

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

(12)

where \( \lambda \) is the forgetting factor, which is usually given in the range of \( 0 < \lambda \leq 1 \), and \( \theta = [p_1(t) p_2(t)]^T \) is the parameter vector to be determined.

Step 2: Calculate derivatives with respect to parameters \( \theta \):

\[
\phi(t) = \frac{dD(t,\theta)}{d\theta}
\]

(13)

Step 3: The parameter update is given by:

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

(14)

and \( P(t) \) is updated as

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

(15)

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

Note that the previous parameter adaptation is realized by a recursive least square method. Some other methods, such as an extended Kalman filter [16], a neural network [15], etc., can be used as well.

V. Experiments

To implement the diagnostic and prognostic algorithm, apart from feature extraction and interpolation, the fault progression model must be initialized. Parameter \( \theta(0) \) is given according to our prior knowledge of the system while \( P(0) \) is given as a large number times an identity matrix. To implement the algorithm, initial parameters are set as:

\[
\begin{bmatrix}
C_D \\
\eta
\end{bmatrix} =
\begin{bmatrix}
0.3 \\
0.5
\end{bmatrix}, \theta(0) =
\begin{bmatrix}
1 \\
1
\end{bmatrix}, P(0) =
\begin{bmatrix}
100 & 0 \\
0 & 100
\end{bmatrix}.
\]

In fault detection, 500 particles are used, while in failure prognosis, to reduce computation time, 30 particles are used. The experimental results on the interpolated data are shown in Figure 7-9. Figure 7 shows the time instant that the fault is detected. Before this time instant, only the fault detection algorithm is running while the failure prognostic algorithm remains inactivated. When the probability of detection reaches a predefined level (90% in the experiment), fault condition is detected and the failure prognostic algorithm is activated immediately. The real-time diagnostic distribution is used as the initial condition for failure prognosis. Since the number of particles for diagnosis and prognosis are different, the diagnostic distribution must be re-sampled to generate 30 particles. To achieve this, the 500
particles of the diagnostic distribution are sorted and a cumulative density function (CDF) is constructed. Then, this CDF is sampled to generate 30 particles and these particles are used as the initial distribution for failure prognosis. Before the implementation of failure prognosis, a hazard value is determined. Usually, this hazard value is a critical fault dimension that endangers the safety of the bearing. In the experiment, this hazard value is given as spall area of 0.06mm². A hazard zone is then defined as a Normal distribution with mean value of the given hazard value and standard deviation of 0.015mm². 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. Figure 8 shows the results at the time instant when failure prognosis is activated.

![Figure 9. Comparison of estimation and actual RUL](image)

The expected value of the estimated RUL pdf, the lower bound of the 95% confidence interval (Just-in-time line [17]) and the actual remaining useful life are shown in Figure 9. It is clear that the estimated RUL is very close to the actual value. Note that the area below the actual remaining useful life is the conservative estimation zone. In practice, and estimate of RUL in the conservative zone is desired because this means the estimated RUL is shorter than the actual RUL. This will lead to early maintenance but will not risk the bearing safety.

VI. CONCLUDING REMARKS

This paper introduces a bearing health monitoring and management methodology that aims to detect incipient failures as early as possible with predefined specifications. The methodology is composed of mathematically rigorous modules including signal processing, feature extraction, fault progression modeling, and fault detection and failure prognosis algorithms. Signal processing aims to remove noise from signal and reduce the variability of the signals while feature extraction is to capture signatures of the fault from measurements. The fault progression model describes the growth of the fault or degradation of bearing’s health state. The establishment of the model and its parameter set determination and adaptation are discussed in detail. The fault detection and failure prognostic algorithms are based on particle filtering with the ability to handle nonlinear systems with non-Gaussian noises. To demonstrate the efficacy of the proposed methodology, experimental results on bearing test data with degradation of its health condition are presented.

REFERENCES