Condition Based Maintenance of Low Speed Rolling Element Bearings using Hidden Markov Model

G. Prakash [1], S. Narasimhan[1], and M. D. Pandey [1]

[1] *University of Waterloo, Waterloo, ON, N2L 3G1, Canada.*
gprakash@uwaterloo.ca
snarasim@uwaterloo.ca
mdpandey@uwaterloo.ca

ABSTRACT

This paper presents an integrated hidden Markov model (HMM) approach to undertake fault diagnosis and maintenance planning for low-speed roller element bearings in a conveyor system. The components studied are relatively long-life components for which run-to-failure data is not available. Furthermore, the large number of these components in a conveyor system makes the individual monitoring of each bearing impractical. In this paper, HMM is employed to overcome both these challenges. For fault diagnosis, a number of bearings varying in age and usage were extracted from the system and tested to develop a baseline HMM model. This data was then used to calculate likelihood probabilities, which were subsequently used to determine the health state of an unknown bearing. For maintenance planning, experimentally determined thresholds from faulty bearings were used in conjunction with simulated degradation paths to parametrize a HMM. This HMM is then used to determine the state duration statistics and subsequently the calculation of residual useful life (RUL) based on bearing vibration data. The RUL distribution is then used for maintenance planning by optimizing the expected cost rate and the results so obtained are compared with the results obtained from a traditional age based replacement policy.

1. INTRODUCTION

Health monitoring (diagnosis), remaining life calculation (prognosis) and maintenance planning (establishing inspection or replacement intervals) of engineering assets are integral to asset management of critical airport infrastructure such as conveyors constituting baggage handling systems (BHS). In practice, these aspects are often decoupled, where fault diagnosis is carried out independently using sensor data (e.g. vibrations), while the latter is undertaken based on reliability principles using life time data of the system or component of interest. While the literature and the suite of tools available for diagnostics —especially using vibration data —are well developed, existing methods for prognosis are applicable only when run-to-failure data (degradation paths) is available. An integrated CBM framework combining all the three aspects: diagnosis, prognosis and maintenance planning is currently lacking for long-life components when such run-to-failure data are unavailable. Widely employed for speech processing (L. Rabiner & Juang, 1986), HMMs offer a versatile CBM framework to unify diagnosis, prognosis and maintenance planning.

HMM's flexible probabilistic structure has resulted in considerable research being carried out for machinery fault diagnosis using vibration measurements (Ertunc, Loparo, & Ocak, 2001; Bunks, McCarthy, & Al-Ani, 2000; Ocak, Loparo, et al., 2001; Mehrabi & Kannatey-Asibu Jr, 2002; J. M. Lee, Kim, Hwang, & Song, 2004; Baruah & Chinnam, 2005; Purushotham, Narayanan, & Prasad, 2005; Bechhoefer, Bernhard, He, & Banerjee, 2006; Nelwamondo, Marwala, & Mahola, 2006; S. Lee, Li, & Ni, 2010). Components such as bearings and cuting tools (Boutros & Liang, 2011), rolling element bearings (Ocak & Loparo, 2005; Purushotham et al., 2005) have been investigated using discrete HMM, and rotor failures using continuous HMM (J. M. Lee et al., 2004). Recently, the authors of the current study (Sadhu, Prakash, & Narasimhan, 2016) applied an improved HMM for fault detection, where raw vibration signals were de-noised using wavelet transform, demodulated using the Teager Kaiser energy operator, and the features were selected using a decision tree. Alshraideh et al. (Alshraideh & Runger, 2014) proposed a general framework to monitor autocorrelated process data (time series data) using HMM and control charts. Despite the volume of literature on the topic of HMMs, a majority of the work is limited to fault diagnosis.

Fault diagnosis and prognosis for machining processes have been pursued using HMM (Baruah & Chinnam, 2005). A

hidden semi Markov model (HSMM, which is a HMM with a temporal structure) was studied for diagnostics and prognostics of pump failure (Dong & He, 2007). Su et al. (Su & Shen, 2013) proposed a novel multi-hidden semi-Markov model to identify degradation and to estimate the remaining useful life of a system, where multiple fused features were used to describe the degradation process. An algorithm for fault diagnosis and RUL estimation using HSMM and HMM for high speed bearings which are short-lived (fast degradation) was recently proposed (Peng & Dong, 2011; Medjaher, Tobon-Mejia, & Zerhouni, 2012). Chen et al. (Chen, Yang, Hu, & Ge, 2011) introduces a bearing fault diagnostics scheme for rotating machinery using multi-sensor mixtured hidden Markov model (MSMHMM). HMM incorporating principal component analysis (PCA) for feature extraction has been proposed for bearing fault prognosis (X. Zhang et al., 2005), where the HMM output (similarity between the current state and the failure state) was called bearing degradation index, which was subsequently extrapolated to estimate the time to reach a predefined failure threshold. In spite of that, the procedure to arrive at these thresholds for replacement was not stated. Moreover, these studies do not address the issue of maintenance planning.

Most studies use log-probabilities or conditional distribution of the state transition for prognostics purposes. For example, some studies utilize decreasing probabilities as a similarity measure (to a healthy bearing) to quantify defect severity (Ocak, Loparo, & Discenzo, 2007). This approach is premised on the assumption that probabilities (similarity) calculated with respect to a healthy HMM model reduce as a bearing deteriorates But, this approach does not yield the RUL distribution. Alternatively, the distribution of time duration for state change conditioned on the current state is predicted (Baruah & Chinnam, 2005). However, for long life components, the time from the current state to the failure state is more important. It has been pointed out (Eker & Camci, 2013) that the duration in any state is a factor that influences the expected future time to be spent in that state, which was ignored in many of the previous works on HMM based prognostics. They presented a discrete-state prognostic method which uses state duration information for RUL estimation. The issue of optimal degradation feature selection for RUL prediction of rolling element bearings was undertaken by zhang et al. (B. Zhang, Zhang, & Xu, 2016). Recently, Wu et. al. (Wu, Tian, & Chen, 2013) demonstrated the use of vibration data collected from bearings to integrate RUL prediction with maintenance planning. Dawid et al. (Dawid, McMillan, & Revie, 2015) presents a review of Markov models, hidden Markov models and partially observable Markov decision processes for maintenance optimization.

The main contributions of this paper are the following: (i) a hybrid approach incorporating multiple degradation paths in conjunction with experimentally obtained thresholds is developed, which allows us to address the case of a slow degrading components where run-to-failure data is unavailable; (ii) fault diagnosis, prognostics and maintenance planning are integrated in an unified framework. For fault diagnosis, several bearings with different usage histories and age are selected from an airport baggage conveyor in operation (i.e., BHS) and tested in a laboratory to develop a baseline HMM model. For maintenance planning, experimentally determined thresholds from faulty bearings were used in conjunction with simulated degradation paths to estimate the parameters of the HMM. This HMM is then used to determine the state duration statistics and hence the calculation of RUL for a given bearing vibration data. This RUL distribution is then used for maintenance planning by optimizing the expected cost rate (ECR) and the results so obtained are compared with the results obtained from a traditional age based replacement policy.

2. BACKGROUND ON HMM

A HMM is a doubly embedded stochastic process with an underlying stochastic process which is not directly observable (hidden) and can be observed only through another stochastic process yielding the sequence of observations (L. R. Rabiner & Juang, 1993). The theory of HMM is based on a Markov Chain (MC). To better understand MC, consider a system described by a set of N distinct states given by $S_1, S_2, ...S_N$. Figure 1 illustrates a Markov process for a system having three $(N = 3)$ states S_1, S_2, S_3. Let the system undergo a

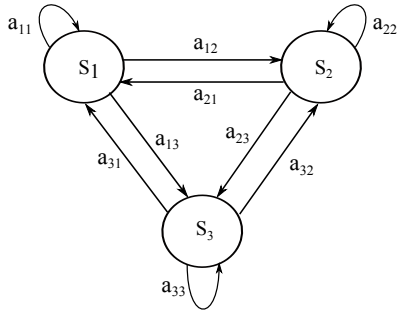


Figure 1. Markov process, ergodic model

transition from one state to another at regularly spaced discrete times, according to the state probabilities (as shown in Figure 1). A full probabilistic description of a Markov process requires specifying the current state as well as the predecessor states. However, for a discrete-time first order Markov process, the probabilistic dependence is truncated to just one, the preceding state. This is also called first order Markov assumption. With this assumption a Markov process can be written as:

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...] = P[q_t = S_j | q_{t-1} = S_i] \quad (1)$$

where, $q_t = S_j$ denotes the state S_j at time t. The transition

matrix $\mathbf{A} = \{a_{ij}\}$, and initial state probability matrix $\boldsymbol{\pi} = \{\pi_i\}$ are given by:

$$\begin{aligned} a_{ij} &= P[q_t = S_j | q_{t-1} = S_i],\ 1 \leq i, j \leq N;\ 1 \leq t \leq T \\ \pi_i &= P[q_0 = S_i] \qquad i = 1, 2, \cdots N \end{aligned} \tag{2}$$

where, a_{ij} is the the probability of transitioning from state S_i to state S_j and π_i is the probability of the process in the i^{th} state at time $t = 0$). The transition matrix $\mathbf{A}$ and the initial state probability $\boldsymbol{\pi}$ together parameterize a Markov model. The probability of observing an observation O_t, given the model λ ($\mathbf{A}$, $\boldsymbol{\pi}$) at a time t, is determined by the joint probability of past and current observations:

$$\begin{aligned} P(O_t|\lambda) &= \prod_{t=1}^{t} P[q_t = S_j, q_{t-1} = S_i] \times P[q_0 = S_i] \\ &= \prod a_{ij} \times \pi_i \end{aligned} \tag{3}$$

In the case of a Markov model, the output of the process is its state, which corresponds directly to a physical, observable event. Nonetheless, in many practical applications, including the case of monitoring bearings using vibration data, an observation is only an indicator of a hidden state of the system. For such cases, HMMs are employed. HMM is an extension of a Markov process and is parameterized by the following elements:

1. N, the number of states in the model.
2. M, the number of distinct observation symbols per state. The individual symbols are denoted by $\boldsymbol{V} = \{v_1, v_2, \cdots v_M\}$.
3. The state-transition probability matrix $\mathbf{A}$, and the initial state probabilities $\boldsymbol{\pi}$, as given by equation (2).
4. The observation symbol probability matrix, $\mathbf{B} = \{b_j(k)\}$, where $b_j(k)$ denotes the probability of emitting a symbol v_k when the system is in state S_j is given by:

$$b_j(k) = P[O_t = v_k | q_t = S_j] \ \ 1 \leq k \leq M \tag{4}$$

Generally, HMM parameters N and M are not known *a priori* and are selected based on the past knowledge of the degradation process or using clustering algorithms (e.g. K-means or Gaussian mixture model (GMM)). Given the three sets of probability measures $\boldsymbol{\lambda}$ = ($\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\pi}$) and model parameters (N, M), there are three basic problems namely: evaluation problem, estimation of optimal state sequence and, the re-estimation of model parameters $\boldsymbol{\lambda}$. The parameter re-estimation and optimal state sequence estimation problems are solved using the *Baum-Wech* and *Viterbi algorithm*, respectively (L. R. Rabiner & Juang, 1993).

2.1 Viterbi algorihm

Viterbi algorithm is used to find the *optimal* hidden state sequence associated with a given observation sequence. To find the best state sequence $\boldsymbol{q} = (q_1 q_2 \cdots q_T)$, for the given observation sequence $\boldsymbol{O} = (O_1 O_2 \cdots O_T)$, we define the quantity $\delta_t(i)$ as follows:

$$\delta_t(i) = \max_{q_1, \cdots, q_{t-1}} P[q_1, q_2, \cdots, q_{t-1}, q_t = i, O_1 O_2 \cdots O_t | \boldsymbol{\lambda}] \tag{5}$$

that is, $\delta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for first t observations and ends in state i. Hence, by induction:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_j(O_{t+1}) \tag{6}$$

To retrieve the the state sequence, we keep the track of the argument that maximized equation (6), for each t and j, which is using this array $\psi_t(j)$. A summary of steps followed is summarized as:

1. Initialization:

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(O_1), \qquad 1 \leq i \leq N \\ \psi_1(i) &= 0. \end{aligned} \tag{7}$$

2. Recursion:

$$\begin{aligned} \delta_t(j) &= \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(O_t) \quad 2 \leq t \leq T \\ \psi_t(j) &= \operatorname*{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \qquad 1 \leq j \leq T \end{aligned} \tag{8}$$

3. Termination:

$$\begin{aligned} P^* &= \max_{1 \leq i \leq N} [\delta_T(i)] \\ q_T^* &= \operatorname*{argmax}_{1 \leq i \leq N} [\delta_T(i)] \end{aligned} \tag{9}$$

4. Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1 \tag{10}$$

2.2 Baum-Welch algorihm

The Baum-Welch algorithm is used to find the unknown parameters $\boldsymbol{\lambda}$ of a HMM. It makes use of the forward-backward algorithm and is named after Leonard E. Baum and Lloyd R. Welch. To describe the iterative procedure for parameters re-estimation, we first define $\xi_t(i, j)$, the probability of being in state i at time t, and state j at time $t+1$, given the model and the observation sequence:

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | \boldsymbol{O}, \boldsymbol{\lambda}) \tag{11}$$

Next, a forward variable $\alpha_t(i)$ is defined as:

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = i | \boldsymbol{\lambda}) \tag{12}$$

that is, the probability of the partial observation sequence $O_1 O_2 \cdots O_t$ (until time t) and state i at time t, given the

model $\boldsymbol{\lambda}$. Similarly, the backward variable $\beta_t(i)$ is given by:

$$\beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T, q_t = i|\boldsymbol{\lambda}) \tag{13}$$

that is, the probability of the partial observation sequence from $t+1$ to the end, given state i at time t and model $\boldsymbol{\lambda}$. From the definitions of the forward and backward variables, $\xi_t(i,j)$ can be written as:

$$\begin{aligned}\xi_t(i,j) &= \frac{P(q_t = i, q_{t+1} = j, \boldsymbol{O}|\boldsymbol{\lambda})}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \\ &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(\boldsymbol{O}|\boldsymbol{\lambda})} \\ &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}\end{aligned} \tag{14}$$

The probability of being in state i at time t i.e., $\gamma_t(i)$, given the entire observation sequence and the model, can be found by summing over j, resulting in:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{15}$$

If we sum $\gamma_t(i)$ over the time index t, we can get the expected number of times that state i is visited (equivalently, the expected number of transitions made from state i). Similarly, summation of $\xi_t(i,j)$ over t (from $t=1$ to $t=T-1$) is the expected number of transitions from state i to the state j. Hence, formulas for re-estimation of HMM parameters $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ can be given as (L. R. Rabiner & Juang, 1993):

$$\bar{\pi}_1 = \text{expected frequency (number of time) in state } i \text{ at time } (t=1) = \gamma_1(i) \tag{16}$$

$$\begin{aligned}\bar{a}_{ij} &= \frac{\text{expected number of transitions from state } i \text{ to state j}}{\text{expected number of transitions from state i}} \\ &= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\end{aligned} \tag{17}$$

$$\begin{aligned}\bar{b}_j(k) &= \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} \\ &= \frac{\sum_{\substack{t=1 \\ s.t. O_t = v_k}}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}\end{aligned} \tag{18}$$

3. THEORETICAL ASPECTS OF RUL ESTIMATION, MAINTENANCE PLANNING AND DEGRADATION MODELING

3.1. RUL estimation and maintenance planning

Consider a bearing with time to failure T that is put into operation at time $t=0$ and still functioning at time t. The probability that the bearing of age t survives an additional interval of length x is

$$R(x|t) = P(T > x+t|T > t) \tag{19}$$

$$= \frac{P(T > x+t)}{P(T > t)} = \frac{R(x+t)}{R(t)} \tag{20}$$

where, $R(x)$ and $R(x|t)$ are the reliability and conditional reliability functions, respectively. The mean remaining useful life of a bearing at age t is given by:

$$\mu(t) = \int_0^\infty R(x|t)dx = \frac{1}{R(t)}\int_t^\infty R(x)dx \tag{21}$$

RUL estimation using HMMs have been undertaken in these references (Tobon-Mejia, Medjaher, Zerhouni, & Tripot, 2011; Chinnam & Baruah, 2009; Dong & He, 2007), where the training data were generated from a single bearing. Alternatively, in this paper, HMMs are trained using vibration data acquired from multiple defective bearings. The key thing to note here is that the run-to-failure histories for these bearings are not available; rather, only the acceleration thresholds corresponding to what were deemed to be faulty bearings during the maintenance process are available. Several degradation paths to these thresholds are simulated (described later). The signal vibration signals are first converted into symbols (see section 6) and subsequently used to train an HMM. In the next step, the estimated parameters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ are used to decode the state (i.e., damage level) sequences and the corresponding stay durations.

For example, Figure 2 illustrates the decoded state sequences and the stay durations for a sample bearing, where D_{ij} denotes the j^{th}stay durations in state $S_i \ \forall \ i,j$. Such state sequence decoding can be performed for multiple run-to-failure bearings. Stay durations in any state can be collected and since these durations follow a Gaussian distribution, which has been verified in Section 3.2, one can estimate the mean and standard deviation for these durations as given below:

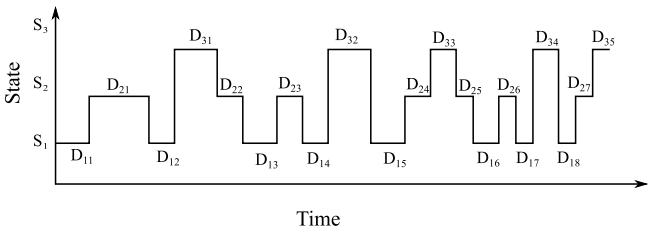


Figure 2. state sequence and stay durations in different state.

$$\mu(D_i) = \frac{\sum_{j=1}^{N} D_{ij}}{N}$$

$$\sigma(D_i) = \sqrt{\frac{\sum_{j=1}^{N}[D_{ij} - \mu(D_i)]^2}{N-1}} \tag{22}$$

where, D_i is the state duration in i^{th} state and N is the total number of times the i^{th} duration is visited corresponding to all bearings under consideration. The Gaussian assumption is verified in Section 3.

Estimation of RUL for a given bearing vibration signal is based on identification of the current state and critical path to reach the failure. The vibration signal is converted into symbols and the state sequence is decoded using the estimated HMM parameters. Here, a new approach for critical path selection is proposed, which is the most probable route by which the component reaches failure from the current state. Amongst various inter-state transitions (from one state to another) along the path, the one with the maximum probability is selected.

For example, Figure 3 shows the case of bearing with states S_1, S_2, S_3, S_4 and transition probabilities, a_{ij} (entries in the matrix $\boldsymbol{A}$) as indicated over the arrows. For a bearing operating in state S_1, the critical path is $S_1 \rightarrow S_2 \rightarrow S_4$. Note that the critical path may or may not be the shortest path. The

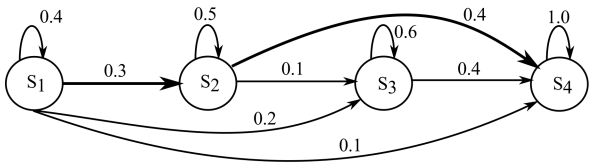


Figure 3. Critical path ($S_1 \rightarrow S_2 \rightarrow S_4$)

temporal parameters (see equation (22)) of each state falling on the critical path are used to estimate the RUL, or failure time (i.e., current time + RUL), which is given by:

$$\text{RUL}_{\text{mean}} = \sum_{\text{i=c}}^{F} \mu(D_i) \tag{23}$$

$$\text{RUL}_{\text{upper}} = \sum_{\text{i=c}}^{F} [\mu(D_i) + n.\sigma(D_i)] \tag{24}$$

$$\text{RUL}_{\text{lower}} = \sum_{\text{i=c}}^{F} [\mu(D_i) - n.\sigma(D_i)] \tag{25}$$

where, c is the current state, F is the failure state and n is the confidence interval coefficient.

The method proposed here is different from the one proposed previously (Tobon-Mejia et al., 2011), where the method for selecting the critical path was based on minimizing the duration to reach the failure state from the current state. This means that all the probabilities in the transition matrix are considered as potential transitions, including reverse transition of states.

Given the RUL estimate from the prognostics phase, the replacement time can be calculated through optimization. One of the widely used preventive replacement policies is age based replacement (ABR), which is based on minimizing the operating cost (Barlow & Hunter, 1960) . A detailed discussion on ABR can be found in this reference (Jardine & Tsang, 2013). Let C_f be the unit cost due to replacement after failure and C_p the unit cost due to preventive replacement (assume $C_f > C_p$). Under this policy, whenever failure occurs, a replacement action is performed and the time is reset to zero and then the component runs for a time t_p, beyond which preventive replacement is done. The optimization problem is to minimize the ECR given by (Jardine & Tsang, 2013):

$$ECR(t_p) = \frac{C_p R(t_p) + C_f[1 - R(t_p)]}{t_p R(t_p) + \int_{-\infty}^{t_p} t f(t) dt} \tag{26}$$

where $f(t)$ and $R(t)$ denote probability density function and the reliability of system, respectively. These parameters are either known through the life time distribution or can be obtained from the prognostics phase (RUL). For example, using the estimated RUL distribution parameters μ and σ (using equation (22) and summing along the critical path), the expressions for $f(t)$ and $R(t_p)$ in equation (26) are given by:

$$f(t) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}; R(t_p) = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{t_p}^{\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{27}$$

As discussed in Section 3.2, the stay duration in any state becomes normally distributed when multiple bearing signatures are used to train the HMM (see Figure 5); hence the RUL is assumed to follow a normal distribution. With equations (26) and (27), the optimal replacement time is obtained by minimizing ECR(t_p), i.e., setting the derivative of equation (26) equal to zero.

3.2. Degradation modeling

Generally, the evolution of a degradation process is monitored over the life of the component through measurements (e.g. vibration), either continuously or periodically. The observed measurements are correlated with the underlying physical degradation process, which can be modeled appropriately. Generally speaking, existing degradation models can be classified into two categories (Lu & Meeker, 1993; Gebraeel, Lawley, Li, & Ryan, 2005; Van Noortwijk, 2009; Elwany & Gebraeel, 2008) : (i) stochastic process models and (ii) random coefficient models, as shown in Figure 4. This classification is general and applies to any degrading system including rolling element bearings. Stochastic processes model the degradation as a cumulative sum of independent and random increments over time, while the latter model deterioration as a linear increase with randomly varying slope (say, Weibull distributed).

For this study, a random coefficient model is adopted as they are suitable to model unit-varying uncertainty. In this model,

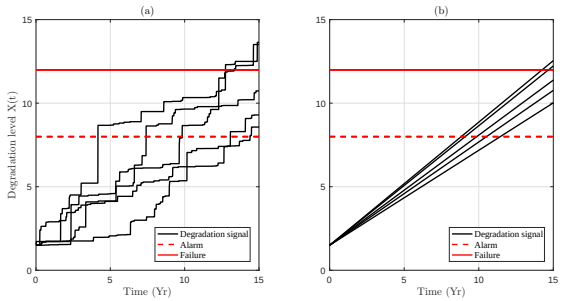


Figure 4. (a) Stochastic degradation (Gamma Process) (b) random coefficient (linear)

a degrading signal is written as:

$$X_k = h(t_k; \phi, \Theta) + \epsilon(t_k) \tag{28}$$

where X_k is the amplitude of the degradation signal monitored using sensors at equally spaced time intervals, $t_k = t \cdot k$, $k = 0, 1...n$. The parameter ϕ captures the constant degradation characteristics over the population, Θ models the unit-varying uncertainty and h is the functional form. The term $\epsilon(t_k)$ is the error due to measurement noise and variability in the signal.

For the BHS bearings, an exponential form is adopted. The choice of an exponential model is justified based on two aspects: (i) the rate of degradation, once a spall is formed, increases with time and, (ii) an abrupt change point is often not found in low-speed components. With an exponential functional form and Brownian error term, the parametric degradation model (see equation (28)) is expressed as (Gebraeel et al., 2005):

$$X_k = \phi + \theta \cdot e^{(\beta \cdot t_k + \epsilon(t_k) - \frac{\sigma^2}{2} \cdot t_k)} \tag{29}$$

where ϕ is a known constant, θ is a lognormal random variable, i.e. $\theta' = \ln(\theta)$ is normal with mean μ_0 and variance σ_0^2, β is a normal random variable, independent of θ with a mean of μ_1 and a variance of σ_1^2, and $\epsilon(t_k)$ is the Brownian motion error with mean 0 and variance $\sigma^2 t_k$. The degradation model in equation (29) can be written in a logarithmic form given by:

$$\ln(X_k - \phi) = \ln\theta + \left(\beta - \frac{\sigma^2}{2}\right) \cdot t_k + \epsilon(t_k) \tag{30}$$

$$\mathrm{L}(t) = \theta' + \beta' \cdot t_k + \epsilon(t_k) \tag{31}$$

where $\beta' = \beta - \frac{\sigma^2}{2}$ and follow a normal distribution with mean μ_1' and variance $\sigma_1'^2$.

Next, we validate the Gaussian distribution for the state durations in the aggregate HMM (HMM trained using multiple degradation signals versus a single degradation path) using simulations. The following parameters $\mu_0 = 3, \sigma_0 = 1.5, \mu_1' = 1, \sigma_1' = 0.5$, and $\sigma = 2$ are used in equation (31) to generate multiple degradation paths. These parameters were selected based on those degradation paths which yield the expected design life of a typical bearing. Two HMMs (3-state) are trained representing a traditional (HMM-1) and an aggregate HMM (HMM-2). For training HMM-1 and HMM-2, one and fifteen simulated degradation signals are considered, respectively. Once the parameters are estimated for the HMMs, the testing is undertaken using degradation signals not used for training. State sequences and stay durations of this signal are decoded with respect to HMM-1 and HMM-2 separately. Stay durations in a particular state (say, state-1) are collected for each case separately and analyzed. Normal quantile-quantile (q-q) plots for these stay durations corresponding to HMM-1 and HMM-2 are shown in Figure 5(a) and Figure 5(b), respectively. Clearly, the stay durations falls approximately on the straight line for the case of an aggregate HMM, thus confirming the normality of durations.

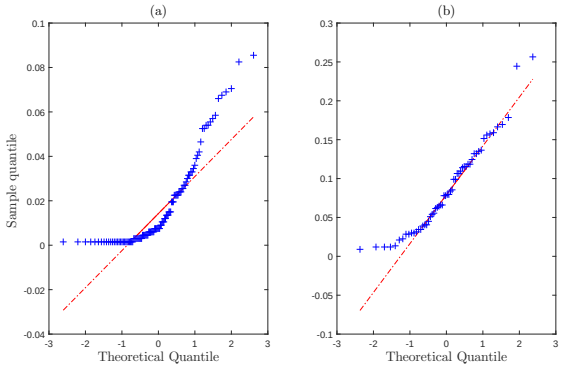


Figure 5. Distribution of stay durations in any state is sensitive to the number of samples used for HMM training. The normal q-q plot for stay durations of one of the decoded states, when (a) one and (b) fifteen degradation signals were used for HMM training.

A key issue in maintenance planning is the setting of thresholds (alarm and failure) based on the amplitude of the degradation signal. The run-to-failure data for individual bearings were not available, however, bearings which were determined to be faulty were made available to the authors. Bearings in their pristine condition were also made available for baseline comparisons. In order to set the thresholds, vibration data was collected from each of these bearings in the prototype conveyor system (details are discussed in Section 5) at the University of Waterloo. Figure 6 shows the q-q plot for the vibration amplitude for three healthy and six faulty bearings. Gaussian distributions were fitted separately to these faulty bearings and the parameters (μ, σ) were estimated. These values were further averaged to give a representative mean, $\bar{\mu}$ (m/s^2) and standard deviation $\bar{\sigma}$ (m/s^2) for faulty bearings. The values, $\bar{\mu} \pm \bar{\sigma}$, were used as alarm limits for preventive maintenance and $\bar{\mu} \pm 2\bar{\sigma}$ as failure limits, which are 8 and 12 m/s^2, respectively. At Pearson airport, Toronto, nearly

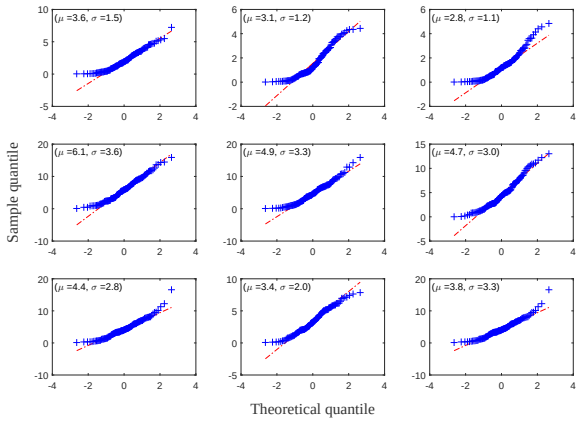


Figure 6. q-q plot for three healthy (first row) and six faulty (second and third row) bearings; experimental acceleration values are shown on the Y-axis and theoretical quantile on the X-axis.

10,000 bearings are in operation within the BHS. Historically, most of the bearing replacements have occurred only on a subset of conveyor sections (details known from historical maintenance logs), which means that the total number of bearings to be monitored is significantly fewer than 10,000. The sample size of the bearings to be monitored can be calculated based on statistical principles and historical bearing failure data. Figure 7 shows a plot of monthly and cumulative bearing replacements undertaken in the last ten years at the airport. The data is negatively skewed (skewness = 1.4) with a standard deviation of $\sigma = 2.7$ years. An approximate estimate of the sample size, n is given by (Asadoorian & Kantarelis, 2005):

$$n = \left[\frac{\sigma \times Z_{\alpha/2}}{E}\right]^2 \tag{32}$$

where, E is the margin of error (i.e., maximum difference between the observed sample mean and the population mean), σ is the population standard deviation and $Z_{\alpha/2}$ is the z-value corresponding to area $\alpha/2$ in the right tail of the standard normal distribution. The available statistics are substituted for population statistics to determine the sample size. Variation of the sample size with error margin (E) and the confidence interval (α) are shown in Table 1. For example, for a 99 percent confidence that the sample mean is within 1 year of the population mean, the sample size $n = 50$. This number can be used either to train a baseline HMM for fault diagnosis or to simulate degradation paths for RUL calculations.

4. OVERALL METHODOLOGY

The proposed method consists of three phases: fault diagnosis, RUL calculation and maintenance planning. The health of a given bearing is determined by calculating the probability of

Table 1. Variation of sample size (n) with error margin (E) and confidence interval (α)

Error Margin (E) (yr.)	Confidence Interval (α) 90%	95%	99%
0.50	80	112	195
0.75	36	50	85
1.00	20	30	50

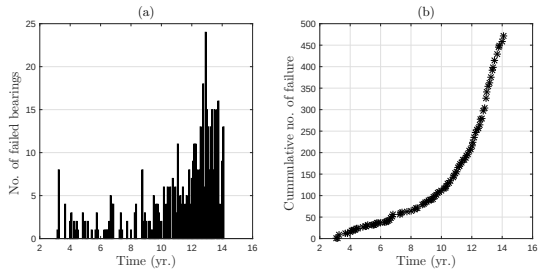


Figure 7. Bearing replacements at Pearson airport (a) monthly (b) cumulative bearing failures.

measurements belonging to a healthy HMM model (based on a threshold). If a bearing is determined to be faulty, an immediate maintenance action is triggered. Else, the RUL pdf is estimated in the second phase, following which a maintenance action is planned based on the estimated RUL in the third phase. Figure 8 shows the algorithm used for fault diagnosis. First, the vibration measurements for healthy bearings at different speeds are collected. Each time series is segmented into several windows of equal length L_w. The optimum window length is determined by maximizing the average kurtosis value $\bar{\mathcal{K}}$, which is given by

$$\bar{\mathcal{K}} = \frac{\sum_{i=1}^{N_w} \mathcal{K}_i}{N_w} = \frac{\sum_{i=1}^{N_w} \sum_{j=1}^{L_w} \frac{(y_{ij}-\bar{y}_i)^4}{\sigma_i^4 L_w}}{N_w} \tag{33}$$

where $\mathcal{K}_i$ is the kurtosis value of ith window, N_w is the total number of window, y_{ij} is the jth data point in ith window, $\bar{y}_i$ is the mean and σ_i is the mean and standard deviation of ith window, respectively. The characteristics of the measurements in each window are captured using condition indicators, namely kurtosis, crest factor, rms value and mean (Večeř, Kreidl, & Šmíd, 2005). This means that the data from each window are represented using a point in a multi-dimensional space, whose co-ordinates are the condition indicators. These points are grouped into clusters using K-means clustering (Duda, Hart, & Stork, 2000) . To train a HMM, a set of alphabetical {H,T,T,H,H,H,T, $\cdots$ }, or numerical {1,2,2,1,1,1,2, $\cdots$ } symbols are generated representing the training sequence. For example, with three clusters, the measurement sequence at a given speed can be coded as a training sequence {1,2,1,2,3,2,1,1, $\cdots$}. Several training sequences are formed by repeating this process at different speeds for healthy bearings and the baseline HMM param-

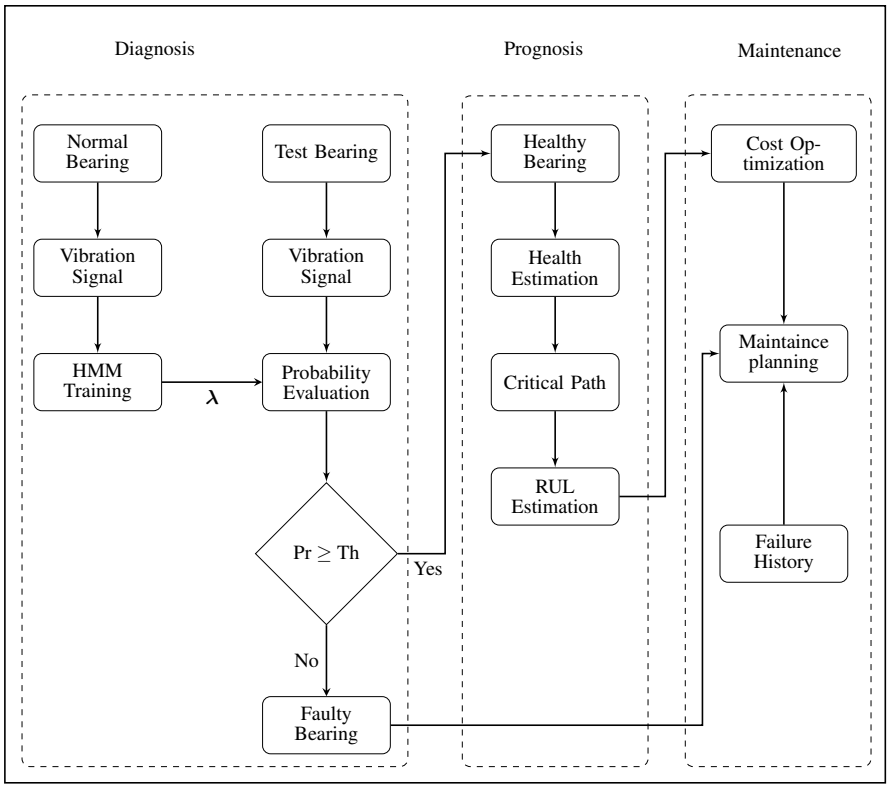


Figure 8. Flowchart showing the key steps of proposed method

eters are estimated. The presence of a fault is determined by setting a threshold calculated using data from multiple healthy bearings, which is the minimum probability amongst all the healthy bearings. Testing involved calculating the probability with respect to baseline HMM and comparing with the threshold. The main steps in the overall methodology are as follows:

Fault diagnosis:

1. Collect vibration measurements from healthy bearings at different speeds.
2. Segment the the signals into multiple windows, evaluate time domain condition indicators (kurtosis and crest factor) for each window and convert them into discrete symbol sequence (described later).
3. Use the *Baum-Welch method* (L. Rabiner & Juang, 1986) to estimate the baseline HMM model with parameters $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.
4. The probability with respect to baseline HMM is evaluated and compared with a threshold. This threshold is the minimum probability amongst a number of healthy bearings experimentally tested with respect to the baseline HMM.

$$\text{Threshold (Th)} = \min\left(P(\mathbf{O}|\boldsymbol{\lambda});\ \forall \text{ normal } \mathbf{O}\right) \quad (34)$$

If the bearing is found to be faulty, immediate replacement is recommended. Else, the RUL is estimated as discussed next.

RUL estimation and maintenance planning steps:

1. Generate simulated run-to-failure vibration measurements and convert them into symbol sequences based on experimentally determined thresholds and estimate the HMM parameters $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.
2. Given $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, use *Viterbi algorithm* (L. Rabiner & Juang, 1986) to decode the state sequence and to calculate the means and standard deviations for the stay durations.
3. For testing, the trained HMM is used to calculate the most recent health state and to identify the critical path from the most recent state to the failure state. The statistics obtained from step 3 are used to calculate the mean and standard deviation for this shortest path and hence the RUL distribution.
4. Given the mean and standard deviation of the RUL, the ECR optimization problem is solved for maintenance planning.

5. EXPERIMENTAL RESULTS

The BHS consists of conveyor units (straight and turn sections) connected in series. Each conveyor is made up of a chassis with two rollers each, one for maintaining the tension

and the other for driving. A gear-motor drives a belt around the two rollers along with a variable frequency drive to vary the operational characteristics such as the belt speed, acceleration, stopping time, etc. A prototype conveyor section was installed at the Structural Engineering Laboratory at the University of Waterloo as shown in Figure 9 and is used for the work described in this paper. This conveyor consists of one straight section (length = 3.83m, width = 1.06m and height = 0.81m) and one turn section (outer curve length = 2.33m and inner curve length = 1.29m). The bearings that guide the motion of conveyor belt between these two sections are shown in Figure 10. Figure 11(a) shows a typical bearing (Dodge, n.d.) supporting the shaft at the junction of straight and turn section. The construction of the bearing, number of balls (n_b), pitch diameter (D_p), ball diameter (D_b) along with the characteristic fault frequencies are provided in Table 2 for reference. The dynamic and static load capacities of this bearing (Dodge-SXR-207-1-7/16) are 22 kN and 15.5 kN, respectively. The L_{10} life for this bearing, which is the life at which ten percent of the bearings can be expected to have failed, equals to 60,000 hour. This estimate is the approximate design life of these bearings. If we assume that the conveyor is running ten hours per day on average, then according to L_{10} the bearing will last for 16.4 years approximately. For the purposes of this study, even though the L_{10} is not explicitly used, this will be used to check the validity of the RUL estimates. The allowable equivalent radial load at 250 RPM (typical conveyor speed) is 3 kN.

Figure 9. Conveyor section located at the University of Waterloo and used for experimental studies.

5.1 Data acquisition

The vibration data was acquired using a tri-axial accelerometer (Dytran, Model 3023A, 10 mV/g sensitivity) mounted on a bearing and the data acquisition system used was manufactured by Datatranslation Inc. (Model DT9837A). Figure 11(b), shows the instrumented bearing on the laboratory test section. Vibration signatures at various locations (3, 6, 9 and 12 o'clock) were collected and analyzed prior to establishing the final 3 o'clock position as the position which results in the best features. The sampling frequency was set to 6 kHz based on the bandwidth of the accelerometer and the features of interest. A tachometer was also employed to measure the rotational speed of the shaft. The conveyor section was operated at speeds ranging from 2.5 - 4.0 Hz, which reflects the typical operating speeds at the airport. Measurements from unloaded and loaded (with four standard baggage specimens weighing 23 Kilograms each) configurations were taken. The accelerations for the loaded configuration were found to be higher in magnitude than their unloaded counterparts.

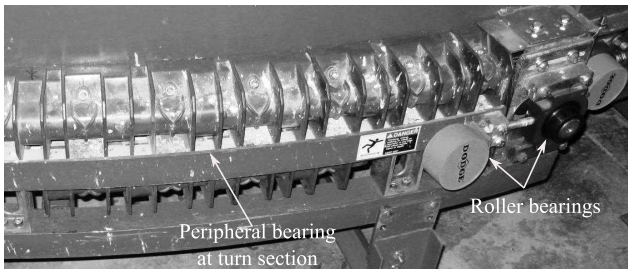
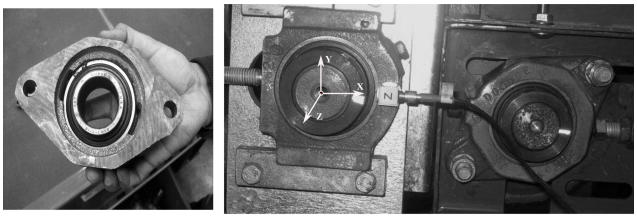


Figure 10. Bearings located at the straight and turn sections.



(a) Installed Dodge bearing (b) Bearing with sensor

Figure 11. Typical Dodge bearings and its instrumentation using a triaxial accelerometer.

Three healthy and six faulty bearings were instrumented and identical tests were conducted on the conveyor section. The faulty bearings were units replaced during the regular maintenance process at airport by the maintenance personnel. Data on both healthy and faulty bearings were acquired at five shaft speeds (172, 191, 212 , 223 and 235 rev/min).

Table 2. Bearing details and characteristics fault frequencies

Geometry			Frequencies (Hz)			
n_b	D_b (in)	D_p (in)	BPFO	BPFI	BSF	FTF
9	0.44	2.136	$3.58\times f_r$	$5.42\times f_r$	$2.34\times f_r$	$0.4\times f_r$

5.2 Analysis and results

Vibration spectra for a healthy and faulty bearing acquired at a speed of 191 RPM (3.18 Hz) are shown in Figure 12. For training, features were extracted from three healthy bearings at speeds 172, 191, 212, 223 and 235 rev/min. To select the optimum window length L_w, the average kurtosis value $\bar{\mathcal{K}}$ for various window lengths is illustrated in Figure 13. It can be seen from the figure that $L_w = 3700$ corresponds to maximum $\bar{\mathcal{K}}$ and hence selected in this study. The number of observation symbols were determined using a Gaussian mixture model (GMM) (Duda et al., 2000). The features were first modeled using GMM and the number of independent components in a GMM constituted the number of observation symbols. The model order (i.e., number of component) was determined using Akaike information criterion (AIC) criteria, which is a measure of the relative quality of statistical models for a given set of data. Mathematically, it is defined as

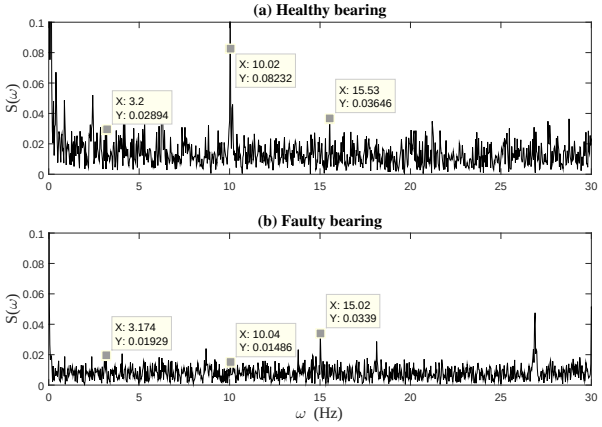


Figure 12. Fourier spectra for (a) healthy and (b) faulty bearing at 191 RPM (3.18 Hz).

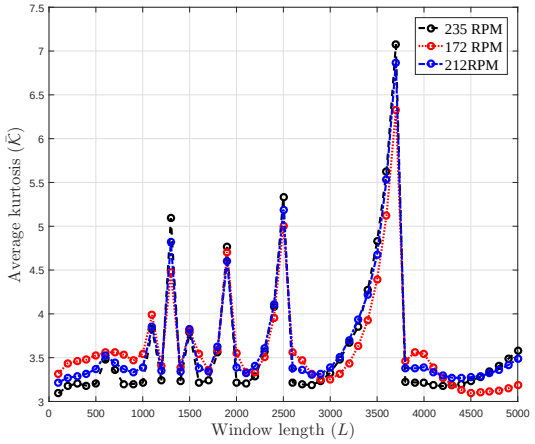


Figure 13. Selection of optimum window length

$$AIC = -2\ln(\hat{L}) + 2\,k \tag{35}$$

where $\hat{L}$ is the maximum value of the likelihood function for the model and k is the number of parameters in the model. The AIC penalizes for the addition of parameters, and thus selects a model that fits well but has a minimum number of parameters (i.e., simplicity and parsimony). Among the several possible models, the one with the lowest AIC value is selected. Table 3 presents the AIC value for various models, when different number of GMM components are considered. Clearly, the three component model is sufficient to represent the extracted features.

Table 3. AIC value with number of GMM component

Number of component	1	2	3	4	5	6
AIC value	540	320	210	212	215	216

Codebook (reference vector) preparation and symbol assignment to the observations were performed next. Training feature vectors from a normal bearing were used in conjunction with K-means clustering to define the three reference vectors (which is the centroid of each cluster) forming the codebook. This code book along with the training vectors are illustrated in Figure 14(a). Assignment of a particular symbol to observations was done by comparing observations with the codebook. For each training vector, the Manhattan distance (Duda et al., 2000) from each reference vector obtained from k-means clustering was used, and the observations were assigned symbols closest to the reference vector. Figure 14(b) shows the assigned symbols (symbols are symbol-1, symbol-2 and symbol-3) for the first hundred observations. For example, observation number one is close to centroid-1 and assigned symbol-1. Similarly, second and third observations are assigned symbol-3 and symbol-1 respectively. The X-axis in Figure 14(a) is not the observation number and an observation can fall anywhere in the plot.

The next step in the training process involves the estimation of the HMM model parameters for a healthy bearing. A set of hundred observation symbols were considered as a observation sequence. For each speed, two such observation sequences were constructed. As an illustration, Figure 16 shows observation sequences generated from 172 RPM (first two rows) and 191 RPM (bottom two rows), respectively. To generate an observation sequence for a given speed, the feature vectors are extracted and compared with the codebook (as prepared in the previous step) and a symbol is assigned. Note the alternation of symbols in the observation sequence in Figure 16, which forms the basis for HMM training. Since measurements for five different speeds were available, the

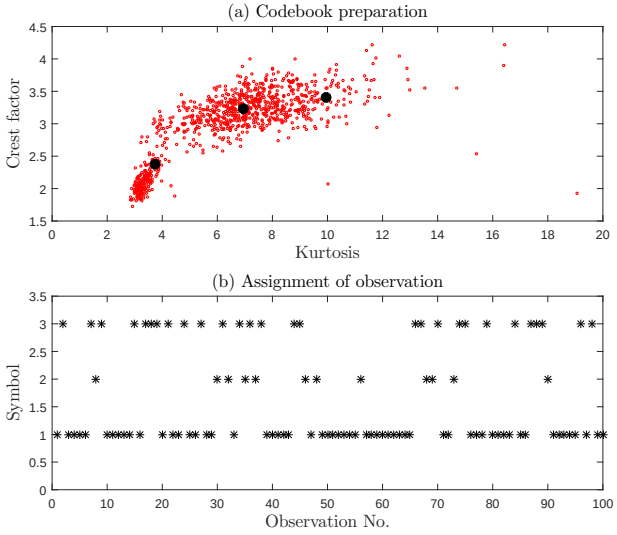


Figure 14. (a) Clusters of condition indicators calculated from vibration measurements for three healthy bearings. The centroid is the reference vector (or code book) which was used for (b) assignment of symbols for the training vectors.

training data consisted of ten observation sequences, where each set contains one hundred observation symbols.

An important issue for HMM training is the number of states, which characterizes the hidden degradation process. Since the number of states are *a priori* unknown, the model was trained assuming various number of states and the optimum number of states was determined from these results. For two and three state HMM, the transition probability **A** and emission probability **B** matrices obtained are given below:
HMM for 2 states:

$$\mathbf{A} = \begin{bmatrix} 0.87 & 0.13 \\ 0.19 & 0.81 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.10 & 0.54 & 0.36 \\ 0.66 & 0.12 & 0.22 \end{bmatrix}$$

HMM for 3 states:

$$\mathbf{A} = \begin{bmatrix} 0.77 & 0.19 & 0.04 \\ 0.07 & 0.67 & 0.26 \\ 0.18 & 0.17 & 0.65 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.20 & 0.68 & 0.12 \\ 0.84 & 0.10 & 0.06 \\ 0.01 & 0.15 & 0.84 \end{bmatrix}$$

Entries in the matrix **A** are the transition probabilities between different states. For example, for the case of three state HMM (S_1, S_2 and S_3), the probability of transition from $S_1 \rightarrow S_2$ is 0.19 and $S_2 \rightarrow S_3$ is 0.04. The maximum probability occurs along the diagonal of the matrix, indicating that the system tends to remains in the same state most of the time. Even though a backward transition, say $S_2 \rightarrow S_1$ is not physically possible, the evaluated transition probability matrix **A** still contains small non-zero values for backward probabilities due to noise and numerical issues (Boutros & Liang, 2011).

Log-probabilities of the training sequences based on two and three state HMMs are given in Table 4. It can be seen from that the probability for the three state HMM is higher than the two state HMM in 90% of the cases tested. This suggests that the given data can be better represented by a three state HMM. The state transition and emission probabilities for the three state HMM are pictorially represented in Figure 15. More complex HMMs with several states would likely to

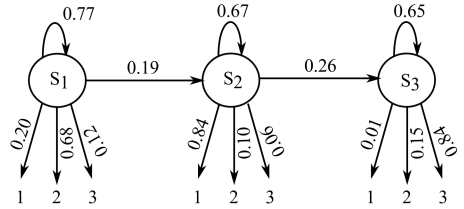


Figure 15. State transition and emission probabilities of different symbols (1,2,3) in a 3-state HMM

capture the underlying deterioration process more efficiently. An adequate number of states can be sought by training various HMM models and comparing their likelihoods with respect to the training data. But, for fault detection in the current set-up, such complex HMMs are not deemed necessary. Also, such HMMs will be computationally expensive. Next, the threshold was decided using the minimum probability obtained for different training sequences for the healthy case in accordance with equation (34). The minimum log-probability for the three state HMM is -106.6, and this is set as the threshold.

The testing phase follows the training phase. Features were extracted and converted into symbols using the codebook. Similar to the healthy case, ten sets of observation sequences were prepared for faulty bearings. The probability of these observations were computed with respect to a normal three state HMM (i.e., P(O|λ)). Log-probabilities for the test sequences are given in Table 4 which shows the resemblance of the observations with the trained HMM model. A log-probability value greater than the threshold signifies greater resemblance to a healthy state. From Table 4 it can be seen that in nine out of ten cases tested (faulty cases), the log-probability is less than the threshold value (-106.6) which confirmed the true state of the bearings used in the experiments.

6. RUL ESTIMATION

Since the run-to-failure data were unavailable, 75 degradation paths simulating a range of bearing usage paths to failure are generated using $\mu_0 = 1.5, \sigma_0 = 0.5, \mu_1' = 1.2, \sigma_1' = 0.5$, and $\sigma = 0.4$ in equation (31). The simulated signals are grouped into three categories: (i) early failure - failure less than 5 years, (ii) middle age failure - failure between 5 to 10 years and (iii) late failure - beyond 10 years, with a probabil-

Table 4. Log-probability of training and test observation sequences

HMM	Log-probabilities of training sequences									
2-State	-86.2	-89.9	-103.3	-98.1	-105.5	-101.4	-109.7	-78.6	-73.3	-68.2
3-State	-78.9	-89.4	-100.1	-92.4	-99.6	-99.5	-106.6	-79.2	-68.4	-69.5
	Log-probabilities of test sequences									
3-State	-120.3	-111.2	-109.1	-115.5	-87.3	-131.4	-112.6	-124.8	-109.4	-111.4

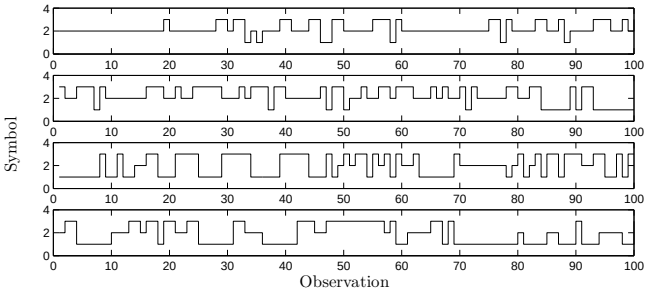


Figure 16. Observation sequences from two speeds, 172 and 191 RPM, used for training

ity of 0.21, 0.31, 0.45, respectively.

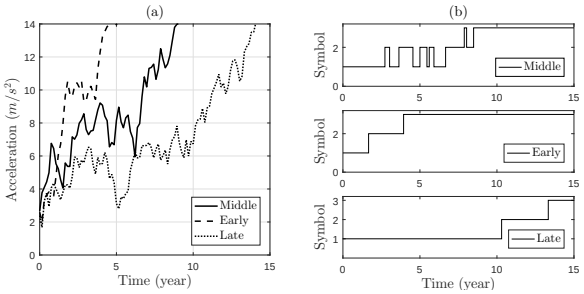


Figure 17. (a) Three BHS bearing degradation paths and (b) corresponding assigned symbols

The acceleration values were converted into symbols {1,2,3}, for accelerations less than 8 m/s^2, between 8 m/s^2 to 12 m/s^2 and greater than 12 m/s^2, respectively. Figure 17 shows three representative signals and the corresponding symbols assigned to them. Subsequently, the HMM parameters are estimated, followed by decoding the state sequences. The estimated time durations for the three states S_1, S_2, S_3 are estimated as:

$$\begin{bmatrix} \mu(D_1) & \sigma(D_1) \\ \mu(D_2) & \sigma(D_2) \\ \mu(D_3) & \sigma(D_3) \end{bmatrix} = \begin{bmatrix} 4.72 & 0.61 \\ 3.01 & 0.37 \\ 2.43 & 0.28 \end{bmatrix} \text{ yr}$$

which is then used to predict the RUL of a test bearing signal. For example, a simulated signal (which has a actual life time of 13.5 years) is used to estimate the RUL at different times. The x-axis represents the point at which inspection is carried out (acceleration data used for analysis) and the y-axis represents the estimated failure time. Figure 18a shows the variation of the estimated failure time and Figure 18b shows the error in its estimation with respect to the actual failure time.

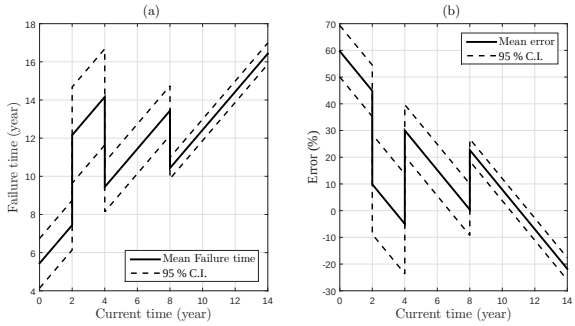


Figure 18. HMM based (a) failure time and (b) estimated error in the failure time .

It is clear that the precision of the estimated failure time increases as the current time approaches the actual failure time. Initially, the mean error in predicting the failure time is about 60%, which decreases to 24% after 5 years and to 8% at the end of 10 years. Hence, as more data becomes available the overall confidence in the RUL estimates becomes higher and hence more useful for predictive maintenance.

7. MAINTENANCE PLANNING

The final step is to optimize the maintenance objective (see equation (26)), given the estimated RUL. To illustrate the traditional age based replacement, we present a numerical example.

Example

Given $C_p = \$10$, $C_f = \$50$, we want to determine the optimal replacement interval of a bearing subjected to age based replacement strategy. Assume that the failures occur according to the normal distribution with a mean (μ) of 10 weeks and a standard deviation (σ) of 2 weeks. With this informa-

tion, equation (26) can be written as

$$C(t_p) = \frac{10 \times R(t_p) + 50 \times [1 - R(t_p)]}{t_p \times R(t_p) + \int_{-\infty}^{t_P} tf(t)dt} \quad (36)$$

Since the failure time is normally distributed failure, the integral $\int_{-\infty}^{t_P} tf(t)dt$ can be simplified as follows:

$$\int_{-\infty}^{t_p} tf(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t_p} t \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right] \quad (37)$$

Applying integration by parts we get

$$\int_{-\infty}^{t_p} tf(t)dt = -\sigma\phi\left(\frac{t_p - \mu}{\sigma}\right) + \mu\Phi\left(\frac{t_p - \mu}{\sigma}\right) \quad (38)$$

where $\phi(t)$ and $\Phi(t)$ are the ordinate and cumulative distribution functions, respectively at t of the standarized normal distribution. For various values of t_p, the corresponding values of $C(t_p)$ are presented in Figure 19, from which it is seen that the optimal replacement age is 6.5 weeks and the corresponding cost is \$1.8.

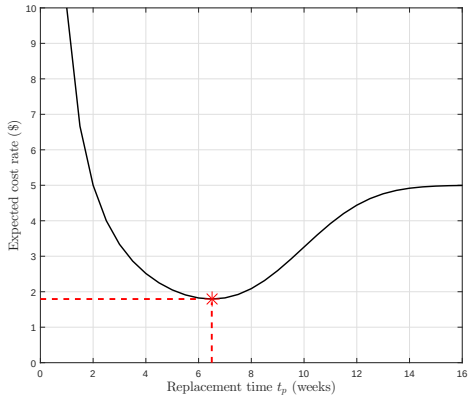


Figure 19. Optimal replacement interval and cost

Now, such calculations is carried out for the simulated bearing acceleration values, where the RUL distribution is obtained from HMM. As discussed previously, the RUL distribution is closer to a normal distribution when multiple bearing signals are used for HMM training. Figure 20(a) shows the estimated RUL for three bearings aged 3, 5 and 7 years (say, for the bearing aged 3 years, the RUL follows a normal distribution $\mathcal{N}(\ 7.2$ yr, 1.2 yr)).

This RUL distribution together with equation (26) are used to frame the maintenance objective function. The preventive replacement cost C_p and the failure replacement cost C_f are assumed (arbitrarily) to be 500\$ and 1000\$, respectively. Variation of expected cost rate with the replacement time (t_p) is illustrated in Figure 20(b), where the optimum replacement time say, for a 3 yr aged bearing is 3.4 yr. Figure 20 also

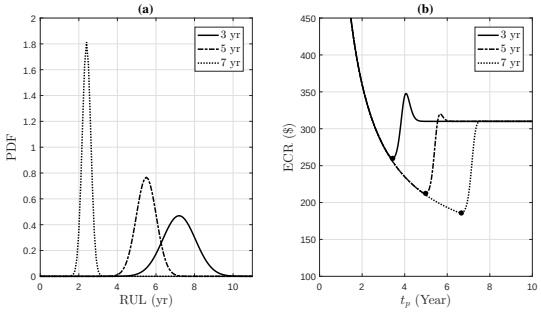


Figure 20. (a) RUL at different time (b) variation of expected cost rate with replacement time.

shows the RUL distribution and the ECR for bearings aged 5 and 7 years. Clearly, the estimates on the replacement time t_p becomes more reliable with increasing time.

Further, the sensitivity analysis for these cost parameters C_p and C_f was carried out and the results are graphically presented in Figure 21. Figure 21(a) shows the variation of ECR with t_p, as a function of the C_p, while keeping C_f fixed. Similarly, the variation of C_f while fixing C_p is shown in Figure 21(b). From the results, it can be seen that the optimal replacement time and cost rate are sensitive to C_p and C_f for the values studied.

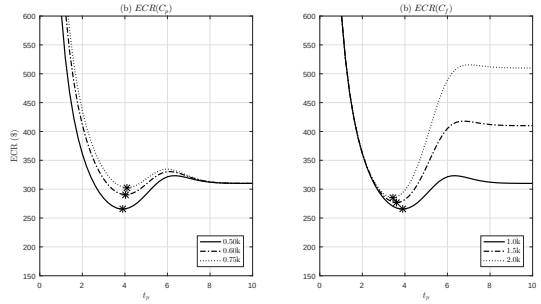


Figure 21. Sensitivity of ECR w.r.t. (a) C_p ($C_f = 1000$\$) and (b) C_f ($C_p = 500$\$.)

Finally, for comparison purposes, the cost rate and replacement interval estimated from the HMM approach are compared with the an age based replacement strategy. To perform such a comparison we first need to estimate the failure time distribution $f(t)$ as in equation (26). A set of 75 degradation paths were simulated, using the same parameters as used in HMM approach. A bearing is considered to be faulty when it reaches the failure threshold of 12 m/s^2 as determined through laboratory experiments. With the aforementioned threshold, over a monitoring period of 15 years, 65 out of 75 bearings were found to have failed, while 10 are censored. Failure time of these bearings is assumed to be Weibull distributed and its probability density function $f(t)$ is esti-

mated, which is subsequently used to estimate the optimum replacement time using equation (26).

Table 5 shows the comparison of t_p and cost rate for these two polices. Note that the age based replacement policy results in a fixed replacement policy i.e., gives only one value of t_p(i.e. 5.5 yr.) based on life time data. On the other hand, the HMM based strategy uses the condition data available up-to the most recent inspection time and updates the replacement time. For example, in Table 5 the results from the HMM approach at the end of 3, 5 and 7 years are given. Clearly, at the end of 7 years, the HMM based policy results in a larger replacement interval and consequently a reduction in the overall cost compared to the traditional age based replacement policy. The effectiveness of HMM based approach can further be verified with respect to the designed L_{10} bearing life, which is 16.4 years as mentioned in section 5. The HMM method predicts bearing replacement time of 13.6 yr. (= 7 + 6.6), when estimated at the end of 7 yr. which is close to the design L_{10} life. Moreover, this prediction will improve further as more data becomes available.

Table 5. HMM based vs age based replacement

Replacement Policy	t_p (yr.)	ECR	(%) $\triangle t_p$	(%) $\triangle$ECR
Age based	5.5	233.1	NA	NA
HMM based (#)	6.6 (7)	186 (7)	+20.0	-20.2
	4.9 (5)	212 (5)	-10.9	-09.1
	3.4 (3)	260 (3)	-38.2	+11.9

[#] Time at which HMM based algorithm is invoked is given in bracket.

8. CONCLUSIONS

It is well known that HMMs are quite useful for bearings health assessment using indirect vibration measurements. However, most existing publications only deal with bearing fault diagnosis, with the exception of very few which deal with the problem of prognosis. In this paper, an integrated HMM framework to undertake fault diagnosis, RUL estimation, and maintenance planning for low-speed rotating components, when failure data is limited or unavailable, is presented. The proposed fault diagnosis algorithm utilizes multiple bearing vibration signals from several bearings at different operating conditions for HMM training and supported by techniques such as GMM, AIC and maximum average kurtosis. Based on the experimental studies performed over a conveyor section in the structural engineering laboratory at the University of Waterloo, it was found that such an approach improves fault detection accuracy. For RUL estimation, a hybrid approach in which experimentally determined thresholds in conjunction with simulated degradation signals is applied to replicate the field situation and address the limited data case. A novel concept of critical path, which is the most probable route by which a system can reach its failure state from the current state is introduced. It is shown that this approach results in better HMM based RUL estimates, which increases with the age of the bearing. Finally, when RUL predictions are integrated to maintenance planning, the results are consistent with the design bearing life, especially in the later stages of operation. Furthermore, the proposed methodology shows cost savings when compared to traditional age based replacement policy.

ACKNOWLEDGEMENTS

The authors would like to thank Greater Toronto Airports Authority and Daifuku Webb for providing the financial support and the experimental test facility to undertake this research study. The authors would also like to thank Natural Sciences Engineering research Council of Canada (Collaborative Research Grants program) and the Ontario Center of Excellence (OCE) for providing matching funds.

REFERENCES

Alshraideh, H., & Runger, G. (2014). Process monitoring using hidden markov models. *Quality and Reliability Engineering International*, *30*(8), 1379–1387.

Asadoorian, M. O., & Kantarelis, D. (2005). *Essentials of inferential statistics*. University Press of America.

Barlow, R., & Hunter, L. (1960). Optimum preventive maintenance policies. *Operations Research*, *8*(1), 90–100.

Baruah, P., & Chinnam, R. B. (2005). HMMs for diagnostics and prognostics in machining processes. *International Journal of Production Research*, *43*(6), 1275–1293.

Bechhoefer, E., Bernhard, A., He, D., & Banerjee, P. (2006). Use of hidden semi-markov models in the prognostics of shaft failure. In *Anuual forum proceedings american helicopter society* (Vol. 62, p. 1330).

Boutros, T., & Liang, M. (2011). Detection and diagnosis of bearing and cutting tool faults using hidden markov models. *Mechanical Systems and Signal Processing*, *25*(6), 2102–2124.

Bunks, C., McCarthy, D., & Al-Ani, T. (2000). Condition-based maintenance of machines using hidden markov models. *Mechanical Systems and Signal Processing*, *14*(4), 597–612.

Chen, Z., Yang, Y., Hu, Z., & Ge, Z. (2011). A new method of bearing fault diagnostics in complex rotating machines using multi-sensor mixtured hidden markov models. In *Proceedings of annual conference of the prognostics and health man-agement society* (pp. 1–6).

Chinnam, R. B., & Baruah, P. (2009). Autonomous diagnostics and prognostics in machining processes through competitive learning-driven hmm-based clustering. *International Journal of Production Research*, *47*(23), 6739–6758.

Dawid, R., McMillan, D., & Revie, M. (2015). Review of markov models for maintenance optimization in the context of offshore wind. , 1–11.

Dodge. (n.d.). *The dodge bearings website.* `http://www.dodge-pt.com/products/bearing/bearinghome/`. ([10-June-2016])

Dong, M., & He, D. (2007). A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing*, *21*(5), 2248–2266.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification (2nd edition)*. Wiley-Interscience.

Eker, O. F., & Camci, F. (2013). State-based prognostics with state duration information. *Quality and Reliability Engineering International*, *29*(4), 465–476.

Elwany, A. H., & Gebraeel, N. Z. (2008). Sensor-driven prognostic models for equipment replacement and spare parts inventory. *IIE Transactions*, *40*(7), 629–639.

Ertunc, H. M., Loparo, K. A., & Ocak, H. (2001). Tool wear condition monitoring in drilling operations using hidden markov models (HMMs). *International Journal of Machine Tools and Manufacture*, *41*(9), 1363–1384.

Gebraeel, N. Z., Lawley, M. A., Li, R., & Ryan, J. K. (2005). Residual-life distributions from component degradation signals: A bayesian approach. *IIE Transactions*, *37*(6), 543–557.

Jardine, A. K., & Tsang, A. H. (2013). *Maintenance, replacement, and reliability: theory and applications*. CRC press.

Lee, J. M., Kim, S.-J., Hwang, Y., & Song, C.-S. (2004). Diagnosis of mechanical fault signals using continuous hidden markov model. *Journal of Sound and Vibration*, *276*(3), 1065–1080.

Lee, S., Li, L., & Ni, J. (2010). Online degradation assessment and adaptive fault detection using modified hidden markov model. *Journal of Manufacturing Science and Engineering*, *132*(2), 021010.

Lu, C. J., & Meeker, W. O. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, *35*(2), 161–174.

Medjaher, K., Tobon-Mejia, D. A., & Zerhouni, N. (2012). Remaining useful life estimation of critical components with application to bearings. *Reliability, IEEE Transactions on*, *61*(2), 292–302.

Mehrabi, M. G., & Kannatey-Asibu Jr, E. (2002). Hidden markov model-based tool wear monitoring in turning. *Journal of Manufacturing Science and Engineering*, *124*(3), 651–658.

Nelwamondo, F. V., Marwala, T., & Mahola, U. (2006). Early classifications of bearing faults using hidden markov models, gaussian mixture models, mel frequency cepstral coefficients and fractals. *International Journal of Innovative Computing, Information and Control*, *2*(6), 1281–1299.

Ocak, H., Loparo, K., et al. (2001). A new bearing fault detection and diagnosis scheme based on hidden markov modeling of vibration signals. In *Acoustics, speech, and signal processing, 2001. proceedings.(icassp'01). 2001 ieee international conference on* (Vol. 5, pp. 3141–3144).

Ocak, H., & Loparo, K. A. (2005). HMM-based fault detection and diagnosis scheme for rolling element bearings. *Journal of Vibration and Acoustics*, *127*(4), 299–306.

Ocak, H., Loparo, K. A., & Discenzo, F. M. (2007). Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: A method for bearing prognostics. *Journal of sound and vibration*, *302*(4), 951–961.

Peng, Y., & Dong, M. (2011). A prognosis method using age-dependent hidden semi-markov model for equipment health prediction. *Mechanical Systems and Signal Processing*, *25*(1), 237–252.

Purushotham, V., Narayanan, S., & Prasad, S. A. (2005). Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden markov model based fault recognition. *Ndt & E International*, *38*(8), 654–664.

Rabiner, L., & Juang, B.-H. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, *3*(1), 4–16.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition* (Vol. 14). PTR Prentice Hall Englewood Cliffs.

Sadhu, A., Prakash, G., & Narasimhan, S. (2016). A hybrid hidden markov model towards fault detection of rotating components. *Journal of Vibration and Control*, 1077546315627934.

Su, C., & Shen, J. (2013). A novel multi-hidden semi-markov model for degradation state identification and remaining useful life estimation. *Quality and Reliability Engineering International*, *29*(8), 1181–1192.

Tobon-Mejia, D., Medjaher, K., Zerhouni, N., & Tripot, G. (2011). Hidden markov models for failure diagnostic and prognostic. In *Prognostics and system health management conference (phm-shenzhen), 2011* (pp. 1–8).

Van Noortwijk, J. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, *94*(1), 2–21.

Večeř, P., Kreidl, M., & Šmíd, R. (2005). Condition indicators for gearbox condition monitoring systems. *Acta Polytechnica*, *45*(6).

Wu, B., Tian, Z., & Chen, M. (2013). Condition-based maintenance optimization using neural network-based health condition prediction. *Quality and Reliability Engineering International*, *29*(8), 1151–1163.

Zhang, B., Zhang, L., & Xu, J. (2016). Degradation feature selection for remaining useful life prediction of rolling element bearings. *Quality and Reliability Engineering International*, *32*(2), 547–554.

Zhang, X., Xu, R., Kwan, C., Liang, S. Y., Xie, Q., & Haynes, L. (2005). An integrated approach to bearing fault diagnostics and prognostics. In *American control conference, 2005. proceedings of the 2005* (pp. 2750–2755).

BIOGRAPHIES

Guru Prakash received his undergraduate degree from the Indian Institute of Technology, Kanpur, India, Master of Applied Science from University of Waterloo, Canada, in 2005 and 2007 respectively. He is currently pursuing his PhD at the University of Waterloo. His research interests is in the area of condition based maintenance planning for long life engineering assets.

Sriram Narasimhan was born in India in 1972. He received his Ph.D. degree from Rice University, USA, in 2005. He joined the University of Waterloo, Canada, in 2006, where he is currently an Associate Professor. In 2014, he was awarded the title of Canada Research Chair (Tier II) in Smart Infrastructure. His main areas of research interest are system identification, control, and vibration based diagnostics. Dr. Narasimhan is a member of the American Society of Civil Engineers and is a licensed engineer in the province of Ontario, Canada.

Mahesh D. Pandey is an active researcher in the areas of risk and reliability analysis and stochastic modeling of engineering problems. He is leading an Industrial Research Chair program sponsored by the Natural Science and Engineering Council of Canada (NSERC) and a consortium of the Canadian nuclear utilities. In the last 10 years under this program, advanced models for reliability and safety analysis of power plant systems have been developed and transferred to the nuclear industry.