Online Tribology Ball Bearing Fault Detection and Identification

B. Ling\textsuperscript{a} and M. M. Khonsari\textsuperscript{b}

\textsuperscript{a}Migma Systems, Inc., 1600 Providence Highway, Walpole, MA 02081
\textsuperscript{b}Department of Mechanical Engineering, Louisiana State University, Baton Rouge, LA 70803
\textsuperscript{a}bling@migmasys.com, \textsuperscript{b}khonsari@me.lsu.edu

ABSTRACT

We present a feasibility analysis for the development of an online ball bearing fault detection and identification system. This system can effectively identify various fault stages related to the evolution of friction within the coated ball bearings. Data are collected from laboratory experiments involving forces, torque and acceleration sensors. To detect the ball bearing faulty stages, we have developed a new bispectrum and entropy analysis methods to capture the faulty transient signals embedded in the measurements. Test results have shown that these methods can detect the small abnormal transient signals associated with the friction evolution. To identify the fault stages, we have further developed a set of stochastic models using hidden Markov model (HMM). Instead of using the discrete sequences, our HMM models can incorporate the feature vectors modeled as Gaussian mixtures. To facilitate online fault identification, we build an HMM model for each fault stage. At each evaluation time, all HMM models are evaluated and the final detection is refined based on individual detections. Test results using laboratory experiment data have shown that our system can identify coated ball bearing faults in near real-time.

Keywords: tribology, feature extraction, fault detection, hidden Markov model, Gaussian mixture

I. INTRODUCTION

Successful operation of many precision machinery used in space instruments require very stringent position accuracy – in the range of microns. These systems must be designed to operate reliably with little or no maintenance and long service-life duration. This presents a formidable challenge to the designer because of the paucity of available information on enabling methods that would ensure proper functioning of these systems. The faults arising in mechanical systems are often linked with bearing faults. In many instances, the accuracy of the instruments and devices used to monitor and control the mechanical system is highly dependent on the dynamic performance of bearings (Goddu 1998).

The most widely studied methods in bearing condition monitoring are based on measurements of vibration, acoustic, or temperature signals. Vibration- and stator-current based methods seem to be some of the most popular in motor health monitoring. When monitoring bearing damage in these motors, the characteristic frequencies of bearing damage are often used to monitor certain frequency components in either vibration or stator current signals (Ilonen 2005). Signals from vibration sensors are usually measured and compared with reference measurements in order to interpret bearing conditions. The methods used to analyze these signals include probabilistic analysis, frequency analysis, wavelet analysis, time-domain analysis, and finite-element analysis. The bearing vibration frequency features and time-domain characteristics are applied to classifiers such as a neural network (Ocak 2001) or HMM (Zhang 2005) to build an automatic bearing fault diagnosis
machine. A survey of available literature reveals, however, that temporal transient features for detection of bearing fault have not been explored.

In Dunbar (2001), a friction fault detection algorithm is designed to monitor friction changes in a mechanical positioning mechanism. It is able to monitor Coulomb and viscous friction separately by an on-line estimation of the friction parameters. The presence of Coulomb friction in pneumatic actuators is a chief obstacle in automated systems that require precision positioning. In Dunbar (2001), instead of directly measuring the friction force for fault detection, the accelerations of the air bearing mass are used to infer the friction force by modeling the dynamic relationship between friction force and acceleration. The fault detection is made possible by monitoring the changing process conditions in the case of wear and excessive side loading in the form of dry friction.

The behavior of friction torque in the precision aerospace instruments of interest is quite unique and different from the classical Columb friction and the Stribeck-type behavior. The quantification of friction involves distinct values known as static, kinetic friction, and viscous friction. In mechanisms that undergo oscillatory motion, friction experiences a discontinuity during the velocity reversal, when the motion traverses through “zero.” This discontinuity renders classical friction models unsuitable for implementation in control applications, in particular when precision positioning is of chief concern (Leonard & Krishnaprasad 1992; Armstrong-Helouvry 1991). The Stribeck friction behavior (Khonsari & Booser 2001) which is frequently used to characterize different regimes of lubrication —i.e., Boundary, Mixed, Elastohydrodynamic, and Hydrodynamic regimes— is also not suited for ultra-low speed motion with velocity reversal.

Recently, Baek & Khonsari (2005) investigated the friction and wear of fluoropolymer coatings in fretting-like oscillatory sliding. The coating was exposed to reciprocating sliding motion of a steel ball in contact with the coating under controlled environment to study the coating life and to assess its contribution to fretting failure which had occurred in a system that utilized these coatings. Extensive sets of experiments revealed that the friction coefficient of the rubber coating with steel ball increases with number of cycles, and the rubber coating is fully worn out after certain number of cycles. Baek & Khonsari (2006) found that the wear debris of polymer coatings acts as a solid lubricant so that the low friction period with the wear debris is very long. But the friction characteristics of the polymer coatings in fretting-like sliding motion are functions of variables such as load, velocity, amplitude and vary with time. Baek & Khonsari (2006) also investigated friction and wear of the polymer coatings with varying temperature and surface roughness in fretting-like sliding motion. Friction coefficient of a fluoropolymer coating shows distinct stages in oscillatory sliding motion.

Under the auspices of NASA STTR Phase I, we have developed a feasibility analysis for the development of a system for fault detection and identification of a ball bearing operating under oscillatory motion. The objectives of the system are to identify, using statistical analysis, various fault modes related to the evolution of friction within the contact in the coated ball bearings. For this purpose, extensive experimental data are collected from laboratory experiments with coated ball bearings. They include measurements of friction force, torque, and acceleration via sensors. To detect the ball bearing fault modes, we have developed a new bispectrum and entropy analysis method to capture the faulty transient signals embedded in the measurements. Test results have shown that this method can detect the small abnormal signals associated with the friction evolution. To identify the fault modes, we have further developed a set of stochastic models using hidden Markov model (HMM). Instead of using the discrete sequences, our HMM models can incorporate the feature vectors modeled as Gaussian mixtures. To facilitate online fault identification, we build an HMM model for each fault mode. At each evaluation time, all HMM models are evaluated and the final detection is refined based on individual detections. The feasibility of software fault detection capability is demonstrated.
This paper is organized as follows: An experimental laboratory system and related data are described in Section 2. The transient torque data analysis is given in Section 3. Then in Section 4, HMM-GMM model used for the detection of ball bearing fault stages is given. And finally, laboratory experimental results are presented in Section 5.

II. EXPERIMENTAL SYSTEM

A series of miniature thrust ball bearings were used for all the tests. Figure 2-1 shows the thrust bearing. The outer diameter and inner diameter are: 12 and 5 mm, respectively. The diameter of the ball is 1.58 mm. A series of tests were initially performed with uncoated bearings and with 4 balls taken out to study the friction characteristic and to establish a baseline. We mainly focus our attention to the performance of coated bearings with 8 balls. A precision friction testing apparatus called Universal MicroTribometer (UMT) manufactured by Center for Tribology Research (CETR) was used in the test as shown in Fig. 2-2.

Each race is attached at the upper and lower part of the fixture, respectively, as shown in Fig. 2-3. A plate spring was used to give a constant load. The lower fixture gives an angular oscillatory motion. The coordinates of the measurement are shown in Fig. 2-4.
The sampling time is 5 ms, the angle of the oscillatory motion is 14.4°, and the rotational speed is 10 rpm. Therefore, the maximum velocity in the middle of the oscillatory motion is about 8.9 mm/sec. The normal load is 42 N. Figure 2-5 shows the X direction force in the beginning of the measurement. The static force in the Y direction is about 1.04 N as shown in Fig. 2-6. These figures represent tests with four (4) balls. Figure 2-7 shows the torque data collected over 25 days.

III. TRANSIENT TORQUE DATA ANALYSIS

The first step in our fault identification method is to calculate the bispectrum from the raw torque data. Let \( \{x(n)\}, n = 0, 1, \ldots, N-1 \), be a discrete time signal. Its bispectrum can be defined as:

\[
B_{31}(f_1, f_2) = \frac{1}{N} X(f_1)X(f_2)\overline{X(f_1 + f_2)}
\]

where \( X(f) \) is the Fourier transform of \( \{x(n)\} \), \( f \) is the normalized frequency.

As bispectrum is a 2-D function, it is difficult to analyze. A 1-D slice of the bispectrum is obtained by freezing one of is two frequency indices. Obviously, there are many types of 1-D slices including diagonal, vertical, or horizontal. The choice of type of slices is problem dependent. Instead of taking 1-D slices, we add bispectra along one frequency index, thus, generating dynamics in the time-frequency domain.
From the bispectrum data, we then calculate the Shannon entropy which is a measure of its spectral distribution of the bispectrum. It is defined as

\[ SE_k = \sum_f |x(f)| \log_2 |x(f)| \]  

(3-2)

where \( k \) is the time index (horizontal axis), \( f \) is the frequency index (vertical axis).

Figure 3-1 shows the raw torque data (top) and its related Shannon entropy of bispectrum. From the torque data, it is difficult to see the fault progress as the ball bearing coating wears out gradually. However, our bispectrum entropy clearly shows the trend (red dashed line) as the entropy gradually increases.

![Figure 3-1: Raw torque data and our estimated Shannon entropy of bispectrum.](image)

The triggers of fault events are estimated by thresholding the bispectrum entropy. Figure 3-2 shows the fault triggers (bottom) for the raw torque data (top). Fault event triggers are the indicators of the presence of fault events. It is clear that the triggers have correctly identified the fault events resulted from the ball bearing coating wear. It is interesting to note that as the coating wears out gradually, the triggers become denser. In other words, fault triggers appear more frequently as coating wear progresses.

![Figure 3-2: Fault triggers (bottom) identified from the raw torque data (top).](image)
From these triggers, we identify three stages: *Normal*, *Fault A*, and *Fault B*, shown in Figure 3-3.

![Figure 3-3: Ball bearing stages estimated from the fault triggers.](image)

We then extract the features from the raw torque data. Our feature space is made of the 2nd moment and kurtosis. The kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. The 2nd moment is a measure of data variation around the statistical mean value. The size of data segment window used for feature calculation is 200 points.

We have extracted 200 feature vectors from each region associated with *Normal*, *Fault A* and *Fault B*. Figure 3-4 shows the locations of these 600 feature vectors (200 from each category). It is clear to see that there are three clusters marked in color RED, BLUE and GREEN. These clusters are not well separated, which is expected since there are transitions among them.

![Figure 3-4: Feature vectors associated with three stages of a ball bearing.](image)

Figure 3-5 shows the correspondence between the clusters and actual regions of the ball bearing stages.
IV. HMM-GMM MODELS FOR BALL BEARING STAGES

To identify the fault modes, we have further developed a set of stochastic models using hidden Markov model (HMM). An HMM is characterized by the following (Rabiner, 1989):

1. The number of states in the model, $N$. We denote the set of states as $S = \{S_1, S_2, \ldots, S_N\}$. Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to sets of states of the model. The state at time $t$ is denoted as $q_t$.

2. The number of distinct observation symbols, $M$. We denote the set of observations as $V = \{v_1, v_2, \ldots, v_M\}$. The observation symbols correspond to the physical output of the system being modeled. The choice of symbols is arbitrary. For example, one can choose the set of observations as $\{1, 2, \ldots, M\}$.

3. The $N \times N$ state transition matrix, $A$, whose $(i, j)$ entry is the probability of a transition from state $i$ to state $j$ defined as follows

   \[ a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i) \]

   \[ a_{ij} \geq 0 \]

   \[ \sum_{j=1}^{N} a_{ij} = 1 \quad \text{for} \quad 1 \leq i, j \leq N. \]

4. The $N \times M$ observation matrix, $B$, whose $(i, k)$ entry is the probability of emitting observation symbol $v_k$ given that the model is in state $i$ defined as follows

   \[ b_{ij} = P(v_k \text{ at } t \mid q_t = S_j) \]

   \[ \text{for} \quad 1 \leq j \leq N, 1 \leq k \leq M. \]

5. The initial state distribution $\pi = \{\pi_i\}$ where
\[ \pi_i = P(q_i = S_i) \]  
for \( 1 \leq i \leq N \).

Given appropriate values of \( N, M, A, B, \) and \( \pi \), the HMM, often represented as \( \lambda = (A, B, \pi) \), can be used as a generator to give an observation sequence

\[ O = O_1 O_2 \ldots O_T \]  

where each observation \( O_t \) is one of the symbols from \( V \), and \( T \) is the number of observations in the sequence.

In most HMM applications, the observation \( O_t \) is assumed to take the discrete values in a set \( O = \{ o_1, o_2, \ldots, o_M \} \). However, for our application, there are no specific discrete observations. In other words, we are dealing with a continuous feature space with certain underlying statistical distribution. Therefore, we must develop an HMM model with output observations drawn values from a feature space. Here we model the distribution of feature vectors as Gaussian Mixture Model.

Given data \( y \) with independent multivariate observations \( y_1, \ldots, y_n \), the likelihood for a mixture model with \( G \) sub-clusters is

\[ L(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G \mid y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \phi_k(y_i \mid \mu_k, \Sigma_k) \]  

where \( \tau_k \) is the probability that an observation belongs to the \( k \)th cluster (\( \tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1 \)), \( \phi_k \) is given as

\[ \phi_k(y_i \mid \mu_k, \Sigma_k) = \frac{\exp\left\{ -\frac{1}{2} (y_i - \mu_k)^T \Sigma_k^{-1} (y_i - \mu_k) \right\}}{\sqrt{\det(2\pi \Sigma_k)}} \]  

The data generated by mixtures of multivariate normal densities are characterized by clusters centered at the means \( \mu_k \) with increased density for points near the mean. The variance matrix \( \Sigma_k \) can be expressed in terms of its eigenvalue decomposition

\[ \Sigma_k = \lambda_k D_k A_k D_k^T \]  

where \( \lambda_k = |\Sigma_k|^{1/d}, D_k \) is the matrix of eigenvectors of \( \Sigma_k \), and \( A_k \) is a diagonal matrix with the normalized eigenvalues of \( \Sigma_k \) on the diagonal in a decreasing order. The parameter \( \lambda_k \) determines the volume of the \( k \)th cluster, \( D_k \) its orientation and \( A_k \) its shape.

As we have seen above, we have identified two fault modes, namely, A and B. Therefore, we have built three independent HMM-GMM models for these modes and NRMAL mode, respectively. These three HMM-GMMs structures are shown in Figure 4-1:
The mode classification can be made based on Bayes’ Theorem. Denote $A_c$ as the event that observation $y$ comes from cluster $c$. We are interested in estimating the posterior probability $P(A_c | y)$. By Bayes’s Theorem,

$$P(A_c | y) = \frac{P(y, A_c)}{P(y)} = \frac{\tau_c f_c(y | \mu_c, \Sigma_c)}{\sum_{k=1}^{G} \tau_k f_k(y | \mu_k, \Sigma_k)}$$  \hspace{1cm} (4-8)

where $c = 1, 2, \ldots, G$. The observation $y$ belongs to cluster $c$ if $P(A_c | y) = \text{Max}\{P(A_k | y), k = 1, 2, \ldots, G\}$.

**V. TEST RESULTS**

We have trained three HMM-GMM models, HMM-GMM$_{\text{normal}}$, HMM-GMM$_{\text{faultA}}$, HMM-GMM$_{\text{faultB}}$. We then process all torque data through these three models. Figure 5-1 shows the transitions of three stages, Normal, Fault A, and Fault B. From this figure, there is no clear indication how these three stages are transitioned.

The main reason why the stage transitions are not conclusive is that there are overlaps among three clusters as indicated in Figure 3-4. Therefore, there are some false transitions. Figure 5-2 shows the transitions over 1000 data points. It is observed that some transitions last for only a few points.
To eliminate false transitions, three sliding window are used to accumulate the number of the same stages occurred in the window. The final stage is determined by choosing the largest number of stage among three windows. Figure 5-3 shows the stage transitions after a sliding window of 5000 points is used. In this figure, value 1 represents NORMAL stage, value 2 FAULT A stage, and value 3 FAULT B stage.

From Figure 5-3, it can be observed that
(a) Three stages are distinctive, indicating that our HMM-GMM models are accurate;
(b) Different stages can be observed during the transitions of stages; and
(c) The trend of stage transition is clear and conclusive.

Figure 5-4 shows the stage transitions together with the raw torque data, bispectrum entropy, and fault event triggers.
From Figure 5-4, it is clear that the HMM-GMM models have accurately predicted the ball bearing stages, namely, Fault A, Fault B, and Normal. These three stages are difficult to identify from the raw torque data. In fact, under the conditions tested the torque data did not show useful trend as ball bearing coating is wearing out. From the density of triggers, one can draw a conclusion that the ball bearing coating is wearing out gradually. However, no information of fault stages can be inferred from the triggers. The HMM-GMM models reported here can be used to identify the fault stages, thus, providing valuable prognosis information for the ball bearings.

**CONCLUSION**

We have developed a feasibility analysis for the development of a system for fault detection and identification of a ball bearing operating under oscillatory motion. Extensive experimental data are collected from laboratory experiments with coated ball bearings. They include measurements of friction force, torque and acceleration via sensors. To detect the ball bearing fault modes, we have developed a new bispectrum, wavegram and entropy analysis method to capture the faulty transient signals embedded in the measurements. To identify the fault modes, we have further developed a set of stochastic models using hidden Markov model (HMM). Instead of using the discrete sequences, our HMM models can incorporate the feature vectors modeled as Gaussian mixtures. To facilitate online fault identification, we build an HMM model for each fault mode. At each evaluation time, all HMM models are evaluated and the final detection is refined based on individual detections. Test results have shown that our HMM-GMM models can be used to identify the fault stages, thus providing valuable prognosis information for the ball bearings.
ACKNOWLEDGEMENT: The work is supported by NASA STTR Phase I funding under contract NND06AA37C. We thank Mr. Ross Hathaway at NASA Dryden Flight Research Center, Edwards, CA, for his valuable suggestions and support.

REFERENCES


