Available online at www.sciencedirect.com
ScienceDirect
Journal of Sound and Vibration 302 (2007) 951–961
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: A method for bearing prognostics

Hasan Ocak[a,*], Kenneth A. Loparo[b], Fred M. Discenzo[c]

[a]*Department of Mechatronics Engineering, Kocaeli University, Kocaeli, Turkey*
[b]*Department of Electrical Engineering & Computer Science, Case Western Reserve University, Cleveland, OH, USA*
[c]*Diagnostics & Sensors Rockwell Automation, Advanced Technology, Mayfield Hts., OH, USA*

Received 31 January 2006; received in revised form 5 January 2007; accepted 8 January 2007
Available online 20 February 2007

Abstract

Bearings are common and vital elements in rotating machinery. By tracking the condition of a bearing, unscheduled machinery outages and costly damage caused by a bearing failure can be avoided. In this paper, we developed a new scheme based on wavelet packet decomposition and hidden Markov modeling (HMM) for tracking the severity of bearing faults. In this scheme, vibration signals were decomposed into wavelet packets and the node energies of the decomposition tree were used as features. Based on the features extracted from normal bearing vibration signals, an HMM was trained to model the normal bearing operating condition. The probabilities of this HMM were then used to track the condition of the bearing. Experimental data collected from a bearing accelerated life test showed that unlike many of the other commonly used trend parameters whose distinguishing features diminished to normal bearing-like levels as the damage grew, the probabilities of the normal bearing HMM kept decreasing as the bearing damage progressed toward bearing failure. As the bearing approached the end of its life (10% of remaining life), the HMM probabilities dropped dramatically signaling severe damage and imminent bearing failure.

1. Introduction

The performance of rotating machinery is very closely related to the condition or the health of the bearings in the system. For example, vibration caused by faulty bearings can significantly degrade the performance of a motor assembly causing downtime and economic loss. Motor failures are often linked to bearing failure. Therefore, bearing condition monitoring can be very cost effective and can reduce the maintenance downtime by providing an advance warning system that allows for the scheduling of timely corrective and repair actions.

Currently, the calculation of the fatigue life of a bearing is mostly based upon the ANSI/AFBMA Standard life rating formula given by

$$L_n = a_1 a_2 a_3 \left(\frac{C}{P}\right)^p, \tag{1}$$

*Corresponding author. Tel.: +905057515625.
E-mail address: hocak@kou.edu.tr (H. Ocak).

0022-460X/$ - see front matter
doi:10.1016/j.jsv.2007.01.001

where L_n, a_1, a_2, a_3, C, P are the rolling contact fatigue life in revolutions $\times 10^6$ corresponding to the reliability level $(100-n)$, reliability factor, material factor, lubrication factor, basic load rating of the bearing and the equivalent load applied to the bearing, respectively. The exponent p is 3 for ball bearings and 10/3 for roller bearings. The life calculations performed using Eq. (1) has its origin in the fatigue life theory of Lundberg and Palmgren [1]. However, the actual life of a bearing can differ significantly from its expected life due to real life operating conditions. Therefore, instead of relying on the statistical life estimates based on laboratory tests, online condition monitoring using vibration analysis has received considerable attention.

Considerable research has been carried out previously for the development of various techniques for bearing fault detection and diagnosis based on vibration. These techniques, which are briefly discussed below, can be classified into time domain, frequency domain, time–frequency domain, higher-order spectral analysis and neural-network techniques. Analyzing the vibration signals directly in the time domain is one of the simplest and cheapest detection and diagnosis approaches. The analysis can be done either by visually examining portions of the time domain waveform (signal morphology) or by examining some statistical parameters related to the time domain vibration signal. Various time domain statistical parameters have been used in Refs. [2,3] as trend parameters to detect the presence of incipient bearing damage. The most commonly used ones are peak, rms, crest factor, kurtosis, and skewness. Crest factor and kurtosis are independent of the actual magnitude of the vibration level and are measures of the spikiness of the vibration signal. In the early stages of bearing damage, kurtosis and the crest factor increase as the spikiness of the vibration increases. However, as the damage increases, the vibration signal becomes more random and the values for the crest factor and kurtosis reduce to more normal bearing like levels. Thus, the statistical analysis approach based on kurtosis and crest factor lacks the ability to detect the bearing defects at the later stages of their development. This is the most important shortcoming of this approach and limits the applicability of these statistical measures for bearing prognosis.

In the frequency domain approach the dominant frequency components of the vibration signals and their amplitudes are used for trending purposes. As faults occur in one of the rolling element bearing components (inner race, outer race, or rolling elements), the vibration spectrum peaks located at the corresponding bearing defect frequency increase, and well as the defect frequency harmonics associated with the defective component(s). In the neighborhood of each peak there are also sidebands, and the spacing of these sidebands depends on the periodic properties of the loading and mechanical transmission path of the vibration signal [4]. As the damage in the bearing increases, the corresponding amplitudes of the peaks in the power spectrum generally increase. Envelope analysis, originally known as the high-frequency resonance technique, is the most commonly used frequency domain-based technique and is explored, in detail, in Ref. [5]. One of the shortcomings of envelope analysis and other frequency domain approaches is that they require the bearing defect frequencies to be known or pre-estimated, and this requires a speed estimate and detailed information about the bearing geometry. The other shortcoming of this approach is the increasing difficulty in analyzing the vibration spectrum when the signal-to-noise ratio is low, in which case fault-imposed frequencies can be masked by noise and other frequency components, especially in the low frequency range where the defect frequencies are located. Recently, there has been some effort on trying to optimize the extraction of the bearing signals from background masking noise through the use of "spectral kurtosis" [6]. In addition, similar to time domain parameters, defect frequency magnitudes in the demodulated spectrum reduce to normal bearing like levels as the damage increases toward bearing failure. Hence, even though envelope analysis can be used as an effective detection and diagnosis method, its utility for bearing prognostics is limited.

Time–frequency domain techniques use both time and frequency domain information allowing for the investigation of transient features, such as impacts. A number of time–frequency domain techniques have been proposed including the short-time Fourier transform (STFT), the Wigner-Ville distribution (WVD), and the wavelet transform (WT) [7,8]. Higher-order spectra describe the degree of phase correlation among different frequencies present in the signal. Large values of phase correlation among the harmonics of any defect frequency are signs of a defect present in the bearing. Bi-coherence spectra are used in Ref. [9] to derive features that relate to the condition of a bearing. The application of bi- and tri-spectral analysis in condition monitoring is discussed in Ref. [10].

Neural networks are applied to bearing fault detection and diagnosis in Refs. [11,12]. In this approach, the bearing diagnostics problem is viewed as a pattern identification problem. Feature extraction techniques are

used to extract features from the vibration signals. These features are then used for training neural networks, which are then used to match the condition of the bearing with one of the possible operating conditions (normal, fault1, fault2, etc.). In order for this approach to work properly, a feature extraction method must be selected so that the features obtained from different fault classes form clusters that separate from each other in the feature space. The performance of the neural-network approach strongly depends on the feature-extraction method used.

In this work, we use the wavelet packet decomposition (WPD), a time–frequency domain technique, in connection with the hidden Markov modeling (HMM) technique for developing a method for the real-time tracking of bearing health and prognostics. HHM is a state of the art technique for time series modeling and classification and has also been successfully applied to many other fields such as tool wear condition monitoring [13] and bearing diagnosis [14]. In the next section we present the theory on WPD and HMMs.

2. Technical background

2.1. Wavelet packet decomposition (WPD)

WTs are widely applied in many engineering fields for solving various real-life problems. The Fourier transform of a signal contains the frequency content of the signal over the analysis window and, as such, lacks any time domain localization information. In order to achieve time localization information it is necessary for the time window to be short, therefore compromising frequency localization. On the contrary to achieve frequency localization requires a large time analysis window and time localization is compromised. Therein lies the dilemma, sometimes referred to as the "uncertainty principle". The STFT represents a sort of compromise between the time and frequency based views of a signal and contains both time and frequency information. STFT has a limited frequency resolution determined by the size of the analysis window. This frequency resolution is fixed for the entire frequency band.

Contrary to STFT, WT provides a more flexible way of time–frequency representation of a signal by allowing the use of variable sized windows. In WT, long time windows are used to get a finer low-frequency resolution and short time windows are used to get high-frequency information. Thus, WT gives precise frequency information at low frequencies and precise time information at high frequencies. This makes the WT suitable for the analysis of irregular data patterns, such as impulses occurring at various time instances.

In the WPD analysis of a signal, the signal is filtered with both low-pass (LP) and high-pass (HP) filters named as quadratic mirror filters. The cut-off frequency of these filters is one fourth of the sampling frequency of the signal. The LP and the HP filtered signals are referred to the approximation (A) and the detail (D), respectively. Both the approximation and the detail are half the size of the original signal and represent the low frequency and the high-frequency content of the signal. If the same decomposition procedure is applied to the first layer A and D signals, a new layer is obtained that consists of four signals each of which is one fourth of the size of the original signal. These new signals are named as AA (approximation of the approximation), DA (detail of the approximation), AD (approximation of the detail) and DD (detail of the detail). The energy of the signal in any layer is referred to as the node energy that we will later utilize as features of a vibration signal. A third level WPD of a signal is illustrated in Fig. 1. In this representation, the third level signals AAA, DAA, ADA, DDA, AAD, DAD, ADD and DDD represent the frequency content of the original signal within the bands 0–$f_s/16$, $f_s/16$–$2f_s/16$, $2f_s/16$–$3f_s/16$, $3f_s/16$–$4f_s/16$, $4f_s/16$–$5f_s/16$, $6f_s/16$–$7f_s/16$ and $7f_s/16$–$f_s/2$, respectively, where f_s is the sampling rate of the signal. A detailed review of WPD can be found in Ref. [15].

2.2. Hidden Markov models (HMMs)

2.2.1. Markov models

Markov models are useful for the analysis of random processes where the probability of occurrence of an event at a given time depends only on the occurrence of the events at the previous time. The idea of a Markov model is best illustrated by the example cited in Ref. [16]. Assume that the weather can be in one of the three states: sunny, rainy, or cloudy. A Markov model of the weather can be constructed as in Fig. 2. Each state of the model represents an observable event, the weather condition. The transition probability a_{ij} is the

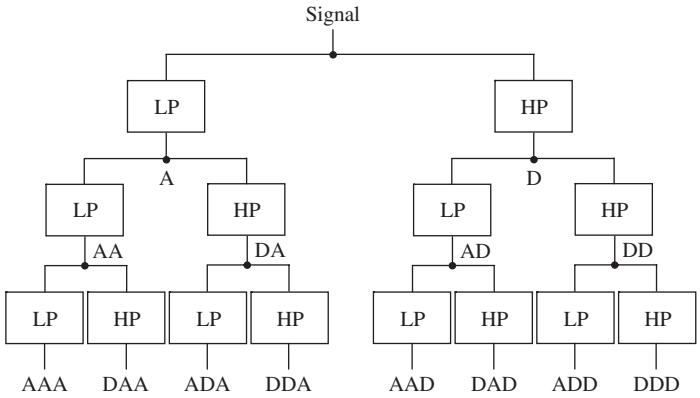

Fig. 1. Third level wavelet packet decomposition of a signal.

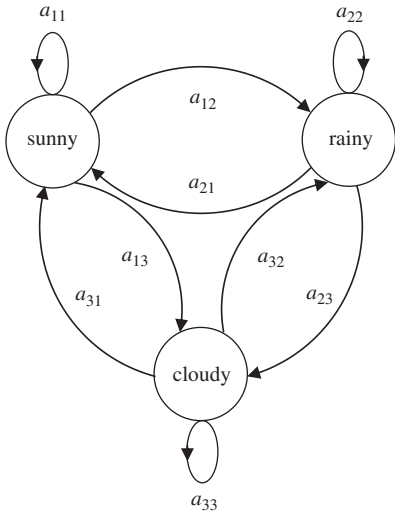

Fig. 2. Three state Markov model of the weather.

probability of tomorrow's weather being in state j provided that today's weather is in state i. This model can be used not only to predict tomorrow's weather but also to calculate the probability of a weather sequence, e.g., sunny–sunny–cloudy–rainy–sunny.

2.2.2. Hidden Markov models (HMMs)

HMMs are the extensions of Markov models to include the case where the observations are probabilistic functions of the states rather than the states themselves. Unlike the states of a Markov model, the states of a HMM do not necessarily represent a physical event. The underlying stochastic process, the state sequence, is not observable, and can only be estimated through another set of stochastic processes that produce the sequence of observations. A three state HMM is illustrated in Fig. 3. In this model, transitions among the states occur only from left to right. This kind of HMM is referred to as a left-to-right HMM. In this example, three distinct observations can be observed in each of the three states with a pre-defined probability.

2.2.3. Elements of a HMM

The following parameters are needed to define a HMM:

- The number of states, N

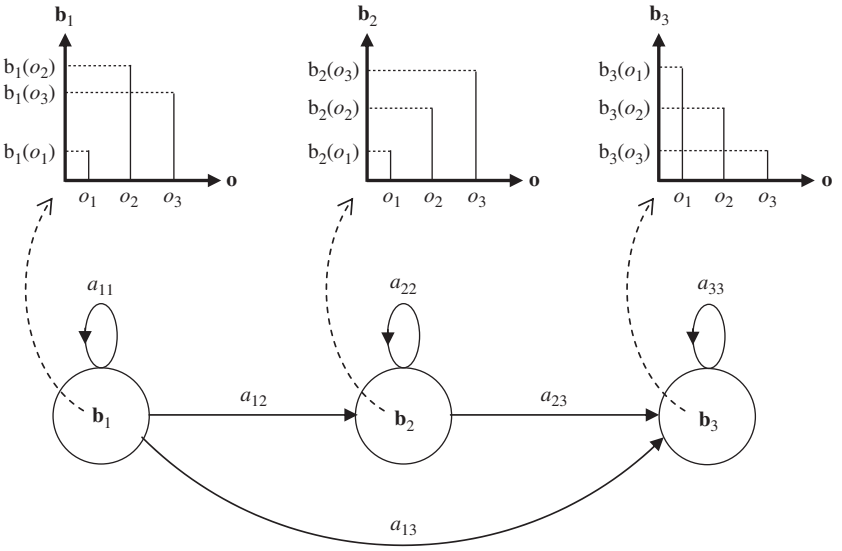

Fig. 3. A three state left-to-right hidden Markov Model.

- The transition probability distribution, $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}, \quad 1 \leqslant i, j \leqslant N, \tag{2}$$

  i.e., the probability of being in state j at time t+1 provided that the state at time t is i.
- Observation probability distribution of each state, $\mathbf{B} = \{b_j(k)\}$, where

$$b_j(k) = p\{o_k | q_t = j\}, \quad 1 \leqslant j \leqslant N, \quad 1 \leqslant k \leqslant M, \tag{3}$$

  where o_k and M denote the kth observation and number of distinct observations, respectively. If the observation is modeled as continuous, a continuous probability density function must be specified for each state.
- The initial state distribution, $\boldsymbol{\pi} = \{\pi\}$, where

$$\pi_i = p\{q_1 = i\}, \quad 1 \leqslant i \leqslant N, \tag{4}$$

  i.e., the probability of the ith state being the initial state.

The compact notation $\boldsymbol{\lambda} = [\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}]$ is used to represent an HMM.

3. Experimental setup for bearing accelerated life testing

A bearing accelerated life test was conducted using the bearing test system shown in Fig. 4. A schematic of the same system showing all components is depicted in Fig. 5. The machine was used to test two 6204 deep groove ball bearings, and is capable of running these bearing at speeds up to 10,000 rev/min and ambient temperatures up to 538 °C. For the accelerated life test, two new 6204-2RS1 SKF bearings were mounted on the shaft. Basic dynamic and static load ratings specified in the catalog for this bearing are 12.7 and 6.55 kN, respectively. 1.51 kN (154 kg) of axial load was applied to these bearings and the operating ambient temperature was set to 127 °C. A PCP 353C65 sensor was stud mounted to the outer housing as shown in Figs. 4 and 5. Data acquisition was established through a National Instruments data acquisition card. The sampling rate was set to 24 kHz. 10 s of data was recorded every hour. On the 50th day of continuous operation, the bearing started to produce extreme noise and vibration. The rotational friction of the rotor caused by the failed bearing increased dramatically causing the system to withdraw very high currents, which made the system shut down automatically.

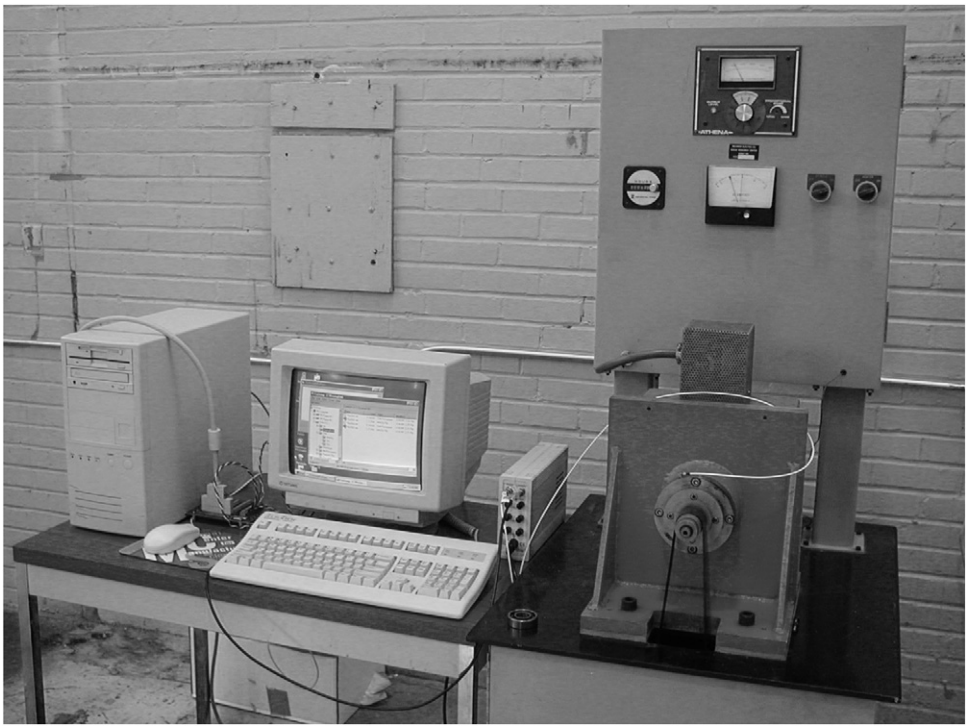

Fig. 4. Bearing test system and data acquisition set-up.

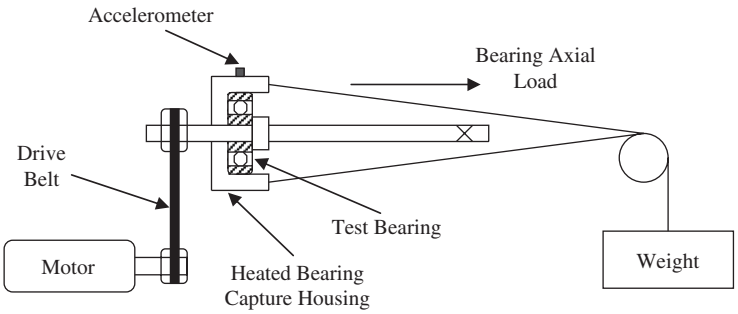

Fig. 5. Schematic of the bearing test system.

4. Methods

4.1. Feature extraction

The spectral content, rather than the vibration amplitude, is the key to determining the bearing condition. Some of the existing frequency domain based prognosis schemes are based upon trending the rms of the vibration signal in certain frequency bands and some are based upon trending the magnitudes of the bearing defect frequencies in the high-frequency demodulated vibration signal. However, by only using the information in certain frequency bands where most but not all of the vibration energy is concentrated, we are losing the information contained in the other frequency regions. The information we are discarding might actually be useful in prognosis. Therefore, instead of relying on certain frequency bands, in this work we utilized the information in the entire frequency spectrum in the feature extraction technique. The vibration signal is divided into several windowed time segments of data of equal length, referred to as epochs. The Nth

level WPD is applied to each of the signal epochs. The feature vector for each epoch consists of the node energies of the Nth level WPD. The feature extraction process is summarized in Fig. 6. Assuming that the vibration signal is divided into T equally spaced epochs, the process will result in a feature matrix of size $T \times 2^N$, where N is the level of the decomposition.

4.2. Diagnosis and severity tracking

A single HMM is trained with features extracted from only normal bearing vibration data. This makes the technique practical to and easy to use. Multiple vibration data collected from a normal bearing throughout a given period of operation where the bearing is known to be normal can be used to train an average model. As the bearing operates in time, feature matrices are extracted from the vibration signals collected from the bearing. Given the feature matrices, the probabilities of the HMM for the normal condition are calculated. The probabilities are then used to assess the condition of the bearing. If the probability is above a pre-determined threshold, then the bearing is classified as being normal. Otherwise, the bearing is classified as being not normal. The lower the probability, the further the current operating condition is from the normal operating characteristics of the bearing. The probabilities tend to decrease over time as the bearing wears. This is contrary to other commonly used trend parameters which capture the severity of wear-induced bearing defects over only a certain operating period and tend to reduce to normal-like levels as the defect becomes increasingly severe. Thus, the HMM based technique is superior to the other bearing related features because the probabilities are monotone and never return to normal-like levels but consistently decrease over time. The experimental results are described in the next section.

5. Results

5.1. Results using traditional trend parameters

First we trended peak-to-peak vibration level, vibration energy and kurtosis values which are directly obtained from the time domain signals. The results are depicted in Figs. 7–9, respectively. The general trend for the peak-to-peak vibration levels showed an increase over time. On the other hand, the general trend for vibration energy showed a decrease during the first 50% of the bearing life and an increase during the remaining 50%. The trend for the kurtosis was just the opposite of the one for the vibration energy. Increasing values of peak-to-peak vibration level signaled bearing wear over time. However, an upcoming bearing failure was forecasted through a big jump in the peak-to-peak vibration level only one day before the actual failure, at the 49th day. Increasing kurtosis values during the first 50% of bearing life also signaled bearing wear over time. However, the decrease in the trend during the remaining 50% of life was misleading and was not predictive of bearing failure. Next, we trended the difference between the defect frequency magnitudes and the background level in the magnitude spectrum of the high frequency demodulated (enveloped) vibration signal.

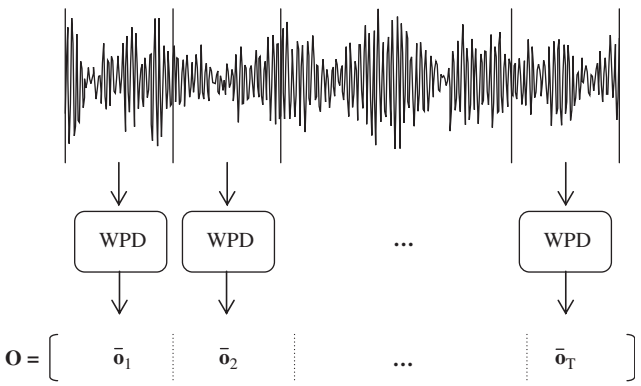

Fig. 6. Vibration feature extraction.

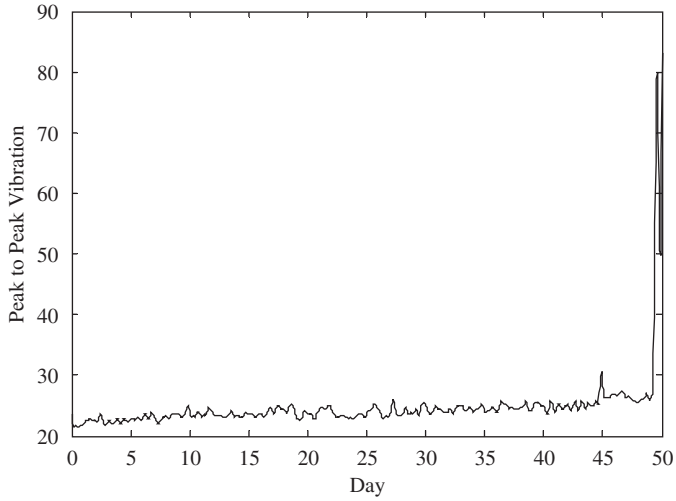

Fig. 7. Peak-to-peak vibration level over time.

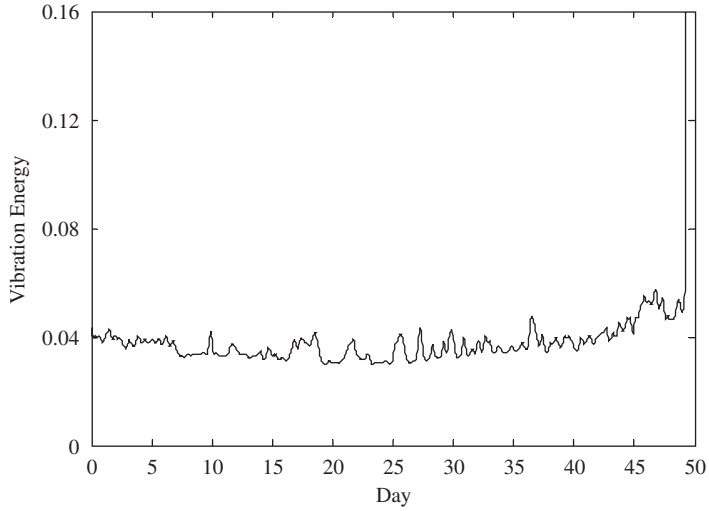

Fig. 8. Vibration energy over time.

Figs. 10–12 show the difference (in dB) between the inner race, outer race and ball defect frequency magnitudes and the background level, respectively. Increasing difference between the magnitude of the spectral component at the inner race defect frequency and the background level signaled an inner race defect with increasing severity over time. The decrease in the trend during the remaining 50% of bearing life period was misleading as was the case with the kurtosis values. The difference between the magnitude of the spectral component at the outer race defect frequency and the background level did not signal any defect on the outer race. Also, the difference between the magnitude of the spectral component at the ball defect frequency and the background level did not signal any significant defect on the rolling elements, either. In conclusion, none of the traditional parameters were able to consistently identify an increasing trend of a severe defect with predictive or prognostic value of an imminent bearing failure except for the peak-to-peak vibration. A large jump in the peak-to-peak vibration level indicated a future failure only one day in advance of the failure event. Besides not being able to indicate a future failure, all the other parameters were also misleading during the last 50% life of the bearing as their signatures returned to normal-like bearing levels.

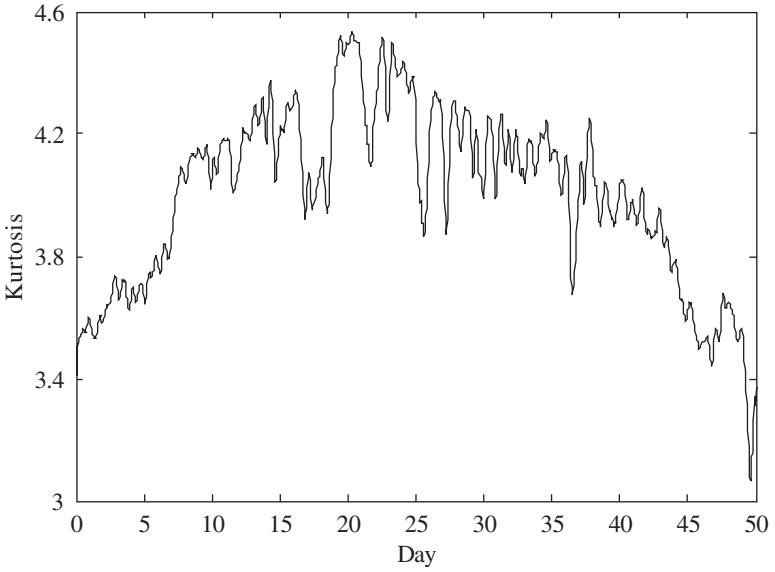

Fig. 9. Kurtosis over time.

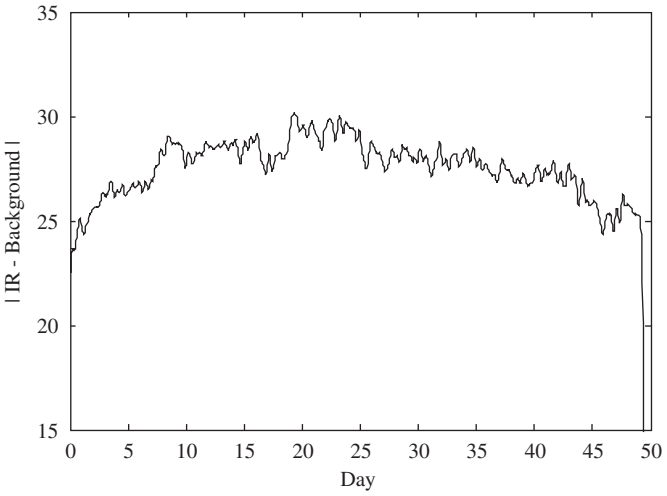

Fig. 10. Difference (in dB) between the inner race frequency magnitude and the background level in the magnitude spectrum of the high frequency demodulated vibration signal over time.

5.2. Results using WPD and HMM based tracking scheme

In the feature extraction process, we used .0427 s (2^{10} data points) epochs with 50% overlap. A 5th level WPD was applied to the segments of the signal in each epoch. The feature vector for each epoch consisted of the node energies of the last level. The model for the normal HMM had three states and three Gaussian distributions were used in the output map for each state. Training data consisted of the vibration signals collected every four hours during the first two days of bearing operation. The logarithm of the probabilities of the feature matrices extracted from the vibration signals given the normal HMM are given in Fig. 13. Decreasing probabilities are indicative of bearing wear over time. The value of the probability can be used as a measure for assessing the defect severity and the trend to bearing failure as these probabilities measure the similarity between the current vibration spectrum and those of a normal bearing. Hence, rather then

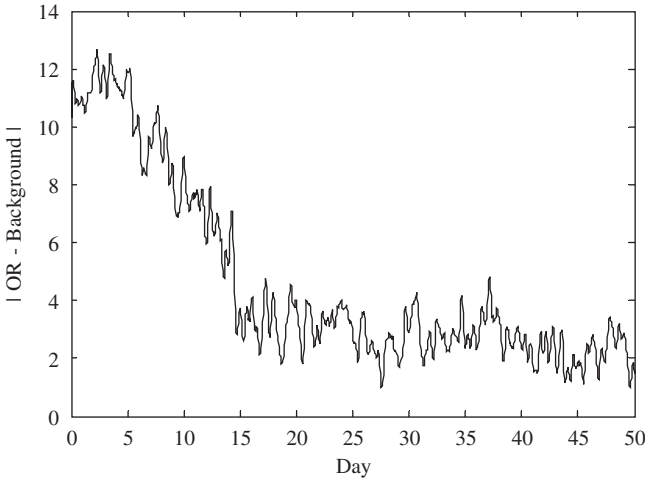

Fig. 11. Difference (in dB) between the outer race frequency magnitude and the background level in the magnitude spectrum of the high frequency demodulated vibration signal over time.

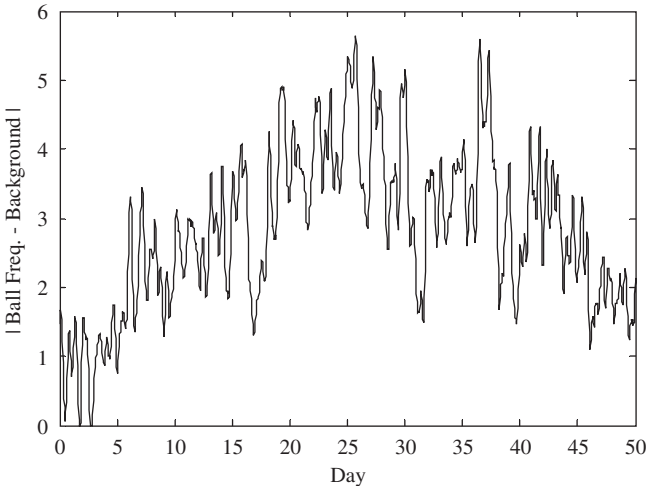

Fig. 12. Difference (in dB) between the ball defect frequency magnitude and the background level in the magnitude spectrum of the high frequency demodulated vibration signal over time.

attempting to track a specific bearing defect and its severity over time we are tracking how different the bearing operation is from normal. Unlike the traditional trend parameters used for prognosis, the HMM probabilities are not misleading during the last 50% of the bearing life. Also superior to other techniques, at the start of the 44th day (the last 10% of remaining useful bearing life), the probabilities dropped dramatically indicating a significant change in bearing behavior and an upcoming bearing failure.

6. Conclusions

A new WPD and HMM based bearing fault tracking scheme was presented in this work. An SKF 6204 bearing was run 50 days until failure. 10 s of data was recorded every hour during the test. An HMM was trained with features extracted from the vibration signals acquired during the first day of operation. To extract features: (1) The signals were divided into equal-sized signal epochs; (2) Nth level WPD was applied to these signal epochs. The feature vector for each window consisted of the node energies of the Nth level. It was shown

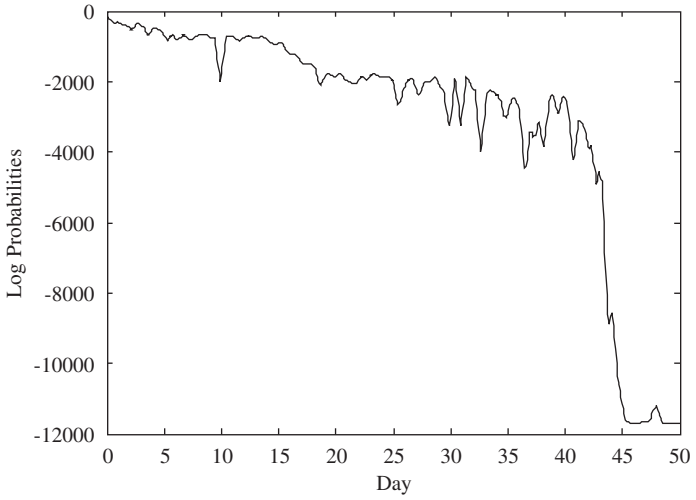

Fig. 13. Log probabilities of the normal HMM over time.

that, the probabilities of the features extracted from the vibration signals and applied to the HMM yielded probabilities that decreased consistently as the bearing wore in time. At the beginning of the final 10% of bearing life, the probabilities dropped dramatically providing a predictive indication of an impending bearing failure.

References

[1] G. Lundberg, A. Palmgren, Dynamic capacity of rolling bearing, *Acta Polytechnica, Mechanical Engineering Series* 2 (4) (1952) 52.
[2] H.R. Martin, F. Honarvar, Application of statistical moments to bearing failure detection, *Applied Acoustics* 44 (1995) 67–77.
[3] R.B.W. Heng, M.J.M. Nor, Statistical analysis of sound and vibration signals for monitoring rolling element bearing condition, *Applied Acoustics* 53 (1) (1998) 211–226.
[4] Y.T. Su, S.J. Lin, On initial fault detection of a tapered roller bearing: frequency domain analysis, *Journal of Sound and Vibration* 155 (1) (1992) 75–84.
[5] P.D. McFadden, J.D. Smith, Vibration monitoring of rolling element bearings by high frequency resonance technique—a review, *Tribology International* 17 (1) (1984) 3–10.
[6] J. Antoni, R.B. Randall, The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines, *Mechanical Systems and Signal Processing* 20 (2) (2006) 308–331.
[7] C.J. Li, J. Ma, Wavelet decomposition of vibrations for detection of bearing localized defects, *NDT & E International* 30 (3) (1997) 143–149.
[8] G.Y. Yen, K.C. Lin, Wavelet packet feature extraction for vibration monitoring, *Proceedings of IEEE International Conference on Control Applications*, IEEE, New York, 1999, pp. 1573–1578.
[9] C.J. Li, J. Ma, B. Hwang, Bearing localized defect detection by bicoherence analysis of vibrations, *ASME Journal of Engineering for Industry* 117 (4) (1995) 625–629.
[10] A.C. McCormick, A.K. Nandi, Bispectral and trispectral features for machine condition diagnosis, *Proceedings of Visual Image Signal Processing*, Vol. 146 (5), IEEE, New York, 1999, pp. 229–234.
[11] M. Subrahmanyam, C. Sujatha, Using neural networks for the diagnosis of localized defects in ball bearings, *Tribology International* 30 (10) (1997) 739–752.
[12] B. Li, G. Godi, M.Y. Chow, Detection of common motor bearing faults using frequency domain vibration signals and a neural network based approach, *Proceedings* of *American Control Conference*, June 1998, pp. 2032–2036.
[13] H.M. Ertunc, K.A. Loparo, H. Ocak, Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs), *International Journal of Machine Tools and Manufacture* 41 (9) (2001) 1363–1384.
[14] H. Ocak, K.A. Loparo, An HMM based fault detection and diagnosis scheme for rolling element bearings, *Journal of Vibration and Acoustics, Transactions of the ASME* 127 (4) (2005) 299–306.
[15] C. Rowden, S. Hall, *Speech Processing*, McGraw-Hill, New York, 1992 (Chapter 5).
[16] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* 77 (2) (1989) 257–286.