Int J Adv Manuf Technol (2006) 30: 738–749
DOI 10.1007/s00170-005-0111-0

ORIGINAL ARTICLE

Ming Dong · David He · Prashant Banerjee · Jonathan Keller

Equipment health diagnosis and prognosis using hidden semi-Markov models

Received: 25 February 2005 / Accepted: 30 May 2005 / Published online: 19 November 2005

Abstract In this paper, the development of hidden semi-Markov models (HSMMs) for equipment health diagnosis and prognosis is presented. An HSMM is constructed by adding a temporal component into the well-defined hidden Markov model (HMM) structures. The HSMM methodology offers two significant advantages over the HMM methodology in equipment health diagnosis and prognosis: (1) it overcomes the modeling limitation of HMM due to the Markov property and therefore improves the power in diagnosis, and (2) it can be directly used for prognosis. The application of the HSMMs to equipment health diagnosis and prognosis is demonstrated with the fault classification application of UH-60A Blackhawk main transmission planetary carriers and prognosis of a hydraulic pump health monitoring application. The effectiveness of the HSMMs is compared with that of the HMMs. The results of the application testing have shown that the HSMMs are capable of identifying the faults under both test cell and on-aircraft conditions while the performance of the HMMs is not comparable with that of the HSMMs. Furthermore, the HSMM-based methodology can be used to estimate the remaining useful life of equipment.

M. Dong (✉)
Department of Industrial Engineering and Management,
Shanghai Jiao Tong University,
1954 Hua Shan Road, Xu-hui District,
200030, Shanghai, PR China
e-mail: mdong@sjtu.edu.cn
Tel.: +86-21-62932115
Fax: +86-21-62932128

D. He · P. Banerjee
Department of Mechanical and Industrial Engineering,
University of Illinois at Chicago,
842 West Taylor Street,
Chicago, IL 60607, USA

J. Keller
U.S. Army RDECOM,
Aviation Engineering Directorate,
Redstone Arsenal, AL 35898, USA

Keywords Hidden semi-Markov model · Condition-based maintenance · Diagnosis · Prognosis

1 Introduction

Condition-based maintenance (CBM) increases system efficiency and availability through elimination of unnecessary maintenance. The economic ramifications of CBM are manifold since they affect labor requirements, replacement part costs, and the logistics of scheduling routine maintenance [4]. A prerequisite to deployment of CBM is effective diagnostics and prognostics.

Recently, a probabilistic approach called hidden Markov models (HMMs) has become increasingly popular and quite effective in some applications such as speech processing and handwritten word recognition. There are two major reasons for this. First, the models have a rich mathematical structure and can form the solid theoretical foundation for use. Second, the models have many successful applications in practice [14]. An added benefit of employing HMMs is the ease of model interpretation in comparison with pure "black box" modeling methods such as artificial neural networks that are often employed in advanced diagnostic models [2]. However, there is an inherent limitation associated with the HMMs. That is, by the assumption of Markovian property, the state duration follows an exponential distribution. For example, an HMM does not provide adequate representation of temporal structure of speech and segmental structure of handwritten words.

Researchers have proposed a number of techniques to address these limitations. Ljolie and Levinson [12] used continuous variable duration HMM in speech recognition. They compared the explicit duration model with the standard HMM. The results showed that the absence of a correct duration model increases the error rate by 50%. The experimental evidence [10, 12, 13] demonstrates that explicit duration models, even if only specifying the longest and the shortest duration allowed for a speech segment, can be beneficial to the performance of the recognizer. As indicated by Chen et al. [5, 6], because of the inherent ambiguity related

to the segmentation process in handwritten words, it is practical to use the variable duration model for the states in a HMM-based HWR system.

There has been limited use of HMMs by the CBM diagnostics community. However, literature does suggest the applicability of these types of models for carrying out diagnostics. Recently, some researchers have applied HMMs in diagnostics of machining processes [1, 3, 8, 9, 15–17]. However, in their applications, only ordinary HMM techniques are adopted. Therefore, the inherent limitation within HMMs as mentioned above still exists in their models. Literature on prognostic methods is extremely limited but the concept has been gaining importance in recent years. Unlike numerous methods available for diagnostics, prognostics is still in its infancy, and literature is yet to present a working model for effective prognostics [2]. Bunks et al. [4] and Baruah and Chinnam [2] first pointed out that HMM-based models could be applied in the area of prognostics in machining processes. However, only standard HMM-based approaches were proposed in their studies.

In this paper, hidden semi-Markov models (HSMMs) for equipment health diagnosis and prognosis are developed. An HSMM is constructed by adding a temporal component into the well-defined HMM structures. The performance evaluation of the developed HSMMs for machinery diagnosis and prognosis is carried out through two real-world application case studies: fault diagnosis of UH-60A Blackhawk main transmission planetary carriers and prognosis of a real hydraulic pump health monitoring application. This paper is organized as follows: Sect. 2 provides a general theoretical background. HSMM-based modeling framework for diagnosis and prognosis is presented in Sect. 3. The basic principle of HSMM-based methodology for diagnosis and prognosis is explained in Sect. 4. Section 5 gives two real-world application case studies of the HSMM-based diagnosis and prognosis method. Finally, Sect. 6 draws conclusions.

2 Theoretical background

2.1 Elements of an hidden Markov model

An HMM represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability function. The generation of a random sequence is then the result of a random walk in the chain and of an observation (also called an emission) at each visit of a state.

An HMM has the following elements [14]:

(1) The state transition probability distribution $A=\{a_{ij}\}$ where

$$a_{ij} = P[s_{t+1} = j | s_t = i] \quad 1 \le i,\ j \le N.$$

(2) The observation probability distribution in state i, $B=\{b_i\ (k)\}$, where

$$b_i(k) = P[v_k | s_t = i] \quad 1 \le i \le N,\ 1 \le k \le M.$$

(3) The initial state distribution $\pi=\{\pi_i\}$ where

$$\pi_i = P[s_1 = i] \quad 1 \le i \le N.$$

(4) N, the number of states in the model, i.e., $1, 2, ..., i, j, ..., N$. Although the states are hidden, there is often some physical signal attached to the states of the model. In this study, we also denote the hidden states as $H=\{h_1, h_2, ..., h_N\}$, and the state at time t as s_t.

(5) M, the number of distinct observations for each state. The observation symbols correspond to the physical output of the system being modeled. The individual observation symbols are denoted as $V=\{v_1, v_2, ..., v_M\}$.

It can be seen that a complete HMM λ requires the specifications of A, B, π and N, M. For convenience, a compact notation is often used in the literature to indicate the complete parameter set of the model:

$$\lambda = (\pi, A, B)$$

2.2 Durational measure of standard HMMs

The durational behavior of an HMM is usually characterized by a durational probability density function (pdf) $P(d)$. For a single state i, the value $P(d)$ is the probability of the event of staying in i for exactly d time units. This event is in fact the joint event of taking the self-loop for $(d-1)$ times and taking the outgoing transition (with probability $1\text{-}a_{ii}$) just once. Given the Markovian assumption, and from probability theory, $P(d)$ is simply the product of all the d probabilities:

$$P_i(d) = a_{ii}^{d-1}(1 - a_{ii}) \tag{1}$$

Here, $P_i(d)$ denotes the probability of staying in state i for exactly d time steps, and a_{ii} is the self-loop probability of state i.

It can be seen that this is a geometrically decaying function of d. It has been argued [18] that this is a source of inaccurate duration modeling with the HMMs since most real-life applications will not obey this function.

2.3 Hidden semi-Markov models (HSMM)

Unlike a state in a standard HMM, a state in an HSMM generates a sequence of observations, as opposed to a single observation in the HMM. Let s_t be the hidden state at time t and O be the observation sequence. Characterization of an HSMM is through its parameters. The parameters for an HSMM are defined as: the initial state distribution (denoted by π), the transition model (denoted by A), state duration

distribution (denoted by D), and the observation model (denoted by B). Thus, an HSMM can be written as $\lambda=(\pi, A, D, B)$.

3 HSMM-based modeling framework for diagnostics and prognostics

For a component, it usually evolves through several distinct health states prior to reaching failure. For example, mechanics of drilling processes suggest that a typical drill bit may go through four health states: good, medium, bad, and worst. In general, for a component, we can identify N distinct sequential states for a failure mechanism. That is, determination of health status of a component: no-defect (i.e., health state 1, denoted by h_1), level-1 defect (denoted by h_2), level-2 defect (denoted by h_3), ..., level-(N-1) defect (denoted by h_N). Here, the level-(N-1) defect means failure. Let d_i be the duration staying at a health state h_i and T be the lifetime of a component. Then, $T=\sum_{i=1}^{N} d_i$.

Unlike a state in a standard HMM, a state in an HSMM model generates a segment of observations, as opposed to a single observation in the HMM. Each health state consists of several single states, which are normal states. Suppose that a health state sequence S has N segments, and let q_n be the time index of the endpoint of the nth segment ($1\le n\le N$). The segments are as follows:

Time units	$1,\cdots,q_1$	$q_1+1,\cdots,q_2$	$\cdots$	$q_{N-1}+1,\cdots,q_N$
Observations	$o_1,\cdots,o_{q_1}$	$o_{q_1+1},\cdots o_{q_2}$	$\cdots$	$o_{q_{N-1}+1},\cdots,o_{q_N}$
Normal states	$s_1,\cdots,s_{q1}$	$s_{q_1+1},\cdots,s_{q_2}$	$\cdots$	$s_{q_{N-1}+1},\cdots,s_{q_N}$
Durations	$d_1=q_1$	$d_2=q_2-q_1$	$\cdots$	$d_N=q_N-q_{N-1}$
Health states	h_1	h_2	$\cdots$	h_N
Segments	1	2	$\cdots$	N

For the nth health state, the observations are $o_{q_{n-1}+1},\cdots,o_{q_n}$, and they have the same normal state label:

$$s_{q_{n-1}+1}=s_{q_{n-1}+2}=\cdots=s_{q_n}\equiv h_n$$

The proposed HSMM-based modeling framework for component diagnostics and prognostics is given in Fig. 1.

4 HSMM-based methodology for diagnosis and prognosis

4.1 Inference procedures

To facilitate the computation in the proposed HSMM-based diagnosis and prognosis framework, in the following, new forward-backward variables are defined and a modified forward-backward algorithm is developed. A dynamic programming scheme is employed for the efficient computation of the inference procedures. To implement the inference procedures, a *forward variable* $\alpha_t(i)$ is defined as the probability of generating $o_1o_2...o_t$ and ending in state i:

$$\alpha_t(i)=P(o_1o_2\cdots o_t, i \quad ends \quad at \quad t|\lambda) \tag{2}$$

Assume that $\alpha_{t-d}(i)$ has been determined, the duration in state j is d. Then $o_1o_2...o_t$ is generated by a state sequence ending at state j if and only if the following conditions are satisfied: (1) $o_1o_2...o_{t-d}$ is generated ending in state i (i.e., $\alpha_{t-d}(i)$); (2) the transition from i to j is chosen (i.e., a_{ij}); (3) the duration in state j is chosen (i.e., $P(d|j)$); and (4) $o_{t-d}o_{t-d+2}...o_t$ is emitted in state j. Summing over all states s and all possible state durations d, we get the recurrence relation [14]:

$$\alpha_t(j)=\sum_{i=1}^{N}\sum_{d=1}^{D}\alpha_{t-d}(i)a_{ij}P(d|j)\prod_{s=t-d+1}^{t}b_j(o_s) \tag{3}$$

where D is the maximum duration within any state.

It can be seen that the probability of O given model λ can be written as:

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i). \tag{4}$$

This is the case since, by definition,

$$\alpha_T(i)=P(o_1o_2\cdots o_T, q_T=i|\lambda) \tag{5}$$

and hence $P(O|\lambda)$ is just the sum of the $\alpha_T(i)$'s.

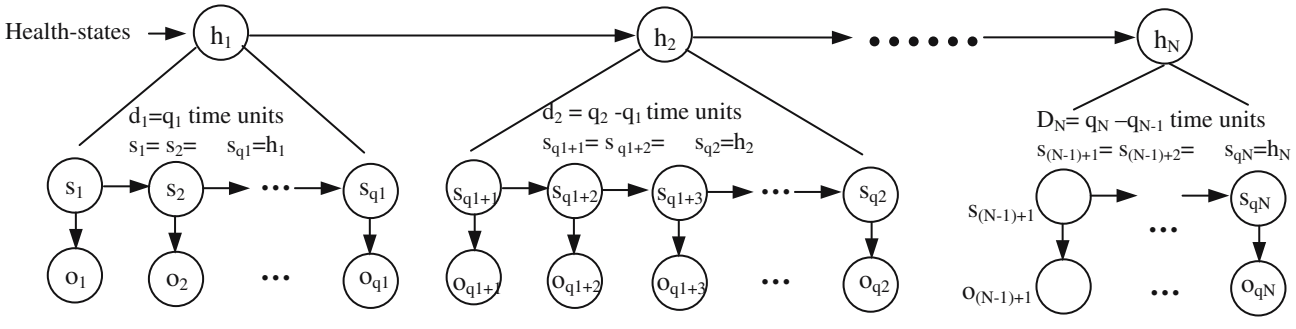


Fig. 1 HSMM-based modeling framework for component diagnostics and prognostics

4.2 Modified forward-backward algorithm for HSMMs

Similar to the forward variable, assume that backward variable $\beta_{t+d}(j)$ has been determined and the duration in state j is d, summing over all states s and all possible state durations d, we have the recurrence relation (see Fig. 2):

$$\beta_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{D} a_{ij} P(d|j) \prod_{s=t+1}^{t+d} b_j(O_s)\, \beta_{t+d}(j) \tag{6}$$

In order to obtain re-estimation formulae for all variables of the HSMM, three more segment-featured forward-backward variables are defined.

$$\begin{aligned} \alpha_{t,t'}(i,j) = & P(o_1 o_2 \cdots o_{t'}, t = q_n, s_t \\ & = i, t' = q_{n+1}, s_{t'} = j|\lambda) \end{aligned} \tag{7}$$

i.e., the probability of the partial observation sequence, $o_1 o_2 ... o_{t'}$, and state i at time t' and state j at time t' ($t'=t+d$).

$$\begin{aligned} \phi_{t,t'}(i,j) = \sum_{d=1}^{D} \Big[& P(d = t' - t|j) \\ & \cdot P\Big(O_{t+1}^{t'} | t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda\Big)\Big] \end{aligned} \tag{8}$$

i.e., the mean value of the probabilities of the system being in state i for d ($d=1,...,D$) time units and then moving to the next state j. Here, $O_{t+1}^{t'} = o_{t+1} o_{t+2} ... o_{t'}$.

$$\xi_{t,t'}(i,j) = P\Big(t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j \Big| O_1^T, \lambda\Big) \tag{9}$$

i.e., the probability of the system being in state i for d ($t' = t+d$) time units and then moving to the next state j, given the observation sequence, $o_1 o_2 ... o_T$. Here, $O_1^T = o_1 o_2 ... o_T$.

$\alpha_{t,t'}(i, j)$ can be described, in terms of $\phi_{t,t'}(i, j)$, as follows:

$$\begin{aligned} \alpha_{t,t'}(i,j) &= P(o_1 o_2 \cdots o_t o_{t+1} \cdots o_{t'}, t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|\lambda) \\ &= P(o_1 o_2 \cdots o_t, t = q_n, s_t = i|\lambda) P\Big(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j \Big| O_1^t, t = q_n, s_t = i, \lambda\Big) \\ &= \alpha_t(i) P\Big(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j | t = q_n, s_t = i, \lambda\Big) \\ &= \alpha_t(i) P(t' = q_{n+1}, s_{t'} = j | t = q_n, s_t = i, \lambda) P\Big(O_{t=1}^{t'} | t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda\Big) \\ &= \alpha_t(i) a_{ij} \sum_{d=1}^{D} P(d = t' - t|j) P\Big(O_{t+1}^{t'} | t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda\Big) \\ &= \alpha_t(i) a_{ij} \phi_{t,t'}(i,j) \end{aligned} \tag{10}$$

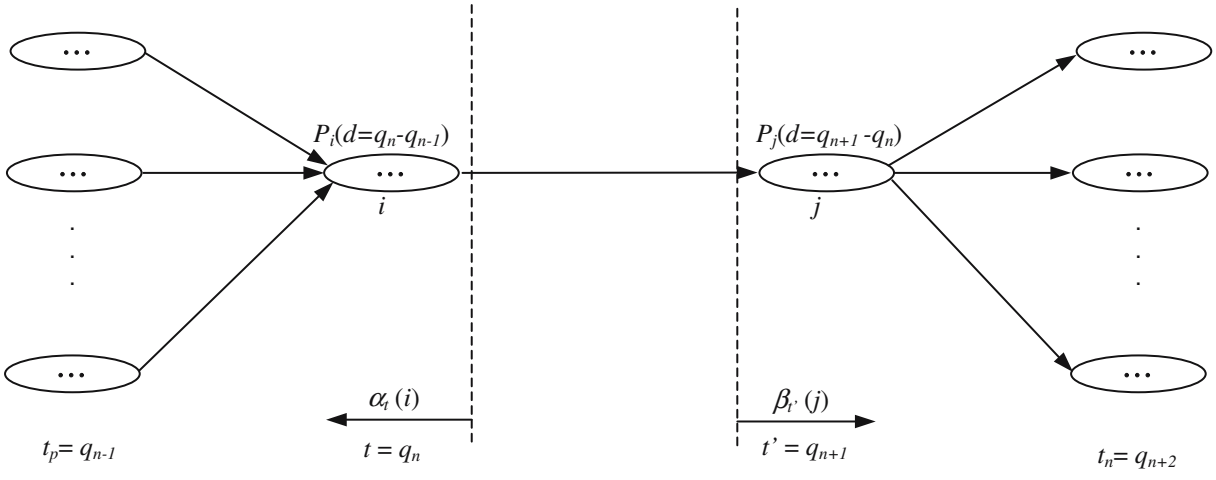


Fig. 2 Illustration of the sequence of operations required for computation of the joint event that the process is in level j at time $t+d$ and level i at time t

The relationship between $\alpha_t(i)$ and $\alpha_{t,t'}(i, j)$ is given in the following:

$$\begin{aligned}\alpha_{t'}(j) &= P(o_1 o_2 \cdots o_{t'}, t' = q_{n+1}, s_{t'} = j|\lambda) \\ &= \sum_{i=1}^{N}\sum_{d=1}^{D} P(d = t' - t|j) P(o_1 o_2 \cdots o_t o_{t+1} \cdots o_{t'} t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j|\lambda) \\ &= \sum_{i=1}^{N}\sum_{d=1}^{D} P(d = t' - t|j)\alpha_{t,t'}(i,j)\end{aligned} \tag{11}$$

From the definitions of the forward-backward variables, we can derive $\xi_{t,t'}(i, j)$ as follows:

$$\xi_{t,t'}(i,j) = \frac{\sum_{d=1}^{D} \alpha_t(i) a_{ij} \phi_{t,t'}(i,j)\beta_{t'}(j)}{\beta_0(i = \text{"}START\text{"})} \tag{12}$$

here, we assume that the state s_0 at time t=0 is a special state "*START*."

The forward-backward algorithm computes the following probabilities:

Forward pass: the forward pass of the algorithm computes $\alpha_t(i)$, $\alpha_{t,t'}(i, j)$ and $\phi_{t,t'}(i, j)$.

Step 1 Initialization (t=0)

$$\alpha_{t=0}(i) = \begin{cases} 1, & if \quad i = \text{"}START\text{"} \\ 0, & otherwise \end{cases}$$

Step 2 Forward recursion (t>0). For t=1, 2,..., T; 1≤i, j≤N; and 1≤d≤D.

$$\begin{aligned}\phi_{t,t'}(i,j) &= \sum_{d=1}^{D} [P(d = t' - t|j) \cdot \\ &\qquad P(O_{t+1}^{t'}|t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda)] \\ \alpha_{t,t'}(i,j) &= \alpha_t(i) a_{ij} \phi_{t,t'}(i,j) \\ \alpha_{t'}(j) &= \sum_{i=1}^{N}\sum_{d=1}^{D} P(d = t' - t|j)\alpha_{t,t'}(i,j)\end{aligned}$$

Backward pass: the backward pass computes $\beta_t(i)$ and $\xi_{t,t'}(i, j)$.

Step 1 Initialization (t=T and 1≤i, j≤N)

$$\beta_T(i) = 1$$

Step 2 Backward recursion (t<T). For t=1, 2, ..., T; 1≤i, j≤N; and 1≤d≤D.

$$\begin{aligned}\beta_t(i) &= \sum_{j=1}^{N}\sum_{d=1}^{D} a_{ij} P(d|j) \prod_{s=t+1}^{t+d} b_j(O_s)\beta_{t+d}(j) \\ &= \sum_{j=1}^{N} a_{ij}\phi_{t,t'}(i,j)B_{t'}(j) \\ \xi_{t,t'}(i,j) &= \sum_{d=1}^{D} \alpha_t(i) a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)/\beta_0(i = \text{"}START\text{"})\end{aligned} \tag{13}$$

Let D_i be the maximum duration for state i. The total computational complexity for the forward-backward algorithm is O(N^2LT), where $L = \sum_{i=1}^{N} D_i$.

4.3 Likelihood equation

Let λ be the set of parameters of a segmental HSMM. For an observation sequence O, and the corresponding state sequence S, the likelihood equation can be written as:

$$\begin{aligned}P(O, S|\lambda) &= \prod_{n=1}^{N} P(s_n|s_{n-1}) \\ &\quad \times \prod_{n=1}^{N} P(d_i = q_n - q_{n-1}|\theta_{d_i}; i = s_n) \\ &\quad \times \prod_{n=1}^{N} P\left(o_{(q_{n-1}, q_n)}|\theta_{f_i}; i = s_n\right)\end{aligned} \tag{14}$$

where θ_{d_i} and θ_{f_i} are the set of parameters for the corresponding distributions.

It is assumed that for each data point o there is a hidden variable s. In an expectation-maximization (EM) algorithm, assume we have λ and estimate the probability that each s occurred in the generation of O. The EM algorithm finds the λ that maximizes the log likelihood $P(O, S|\lambda)$. It

starts from some initial guess $\lambda^{(0)}$ of λ, and then iterates over the following two steps:

E-step: compute the distribution $P(s|\ o,\ \lambda^{(0)})$ for each data point o and the corresponding hidden state s.

M-step: set λ^{new} to the λ that maximizes the expected full log likelihood:

$$\lambda^{new} = \arg\max_{\lambda} \sum_{0} \sum_{s} P\left(s|o, \lambda^{old}\right) \log P\left(o, s|\lambda^{old}\right) \tag{15}$$

In the following, we decompose the expected full log likelihood in the M-step above into three parts, each with its own subset of parameters:

$$\begin{aligned}
&\sum_{o} \sum_{s} P(s|o, \lambda^{old}) \log P(o, s|\lambda^{old}) \\
&= \sum_{o} \sum_{s} P(s|o, \lambda^{old}) \cdot \\
&\quad \left[\sum_{n=1}^{N} \log\ P(s_{q_n}|s_{q_{n-1}}, \lambda^{old}) \right. \\
&\quad + \sum_{n=1}^{N} \log P(d = q_n - q_{n-1}|s_{q_n} = i, \lambda^{old}) \\
&\quad \left. + \sum_{n=1}^{N} \log P(o^{q_n}_{q_{n-1}+1|s_{q_n}} = i, \lambda^{old}) \right] \\
&= \sum_{o} \sum_{s} P(s|o, \lambda^{old}) \sum_{n=1}^{N} \log P(s_{q_n}|s_{q_{n-1}}, \lambda^{old}) \\
&\quad + \sum_{o} \sum_{s} P(s|o, \lambda^{old}) \sum_{n=1}^{N} \log P(d = q_n - q_{n-1}|s_{q_n} = i, \lambda^{old}) \\
&\quad + \sum_{o} \sum_{s} P(s|o, \lambda^{old}) \sum_{n=1}^{N} \log P\left(o^{q_n}_{q_{n-1}+1|s_{q_n}} = i, \lambda^{old}\right)
\end{aligned} \tag{16}$$

4.4 Parameter re-estimation for HSMM-based framework

4.4.1 Initial state distribution

The re-estimation formula for initial state distribution is the probability that state i was the first state, given O.

$$\overline{\pi}_i = \frac{\pi_i \left[\sum_{d=1}^{D} \beta_d(i) P(d|i) b_j\left(O_1^d\right) \right]}{P(O|\lambda)} \tag{17}$$

4.4.2 State transition probabilities

The re-estimation formula of state transition probabilities is the ratio of the expected number of transitions from state i to state j, to the expected number of transitions from state i.

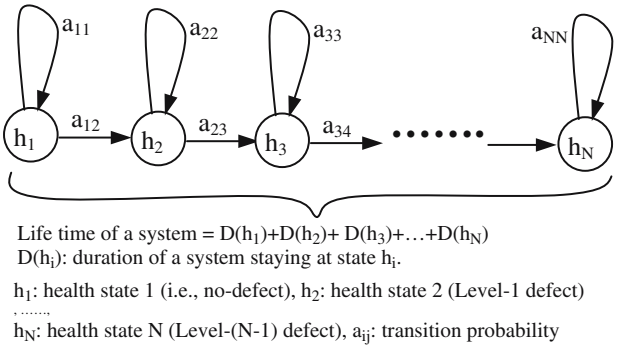


Fig. 3 HSMM-based prognostics

$$\begin{aligned}
\overline{a}_{ij} &= \frac{\sum_{t=1}^{T} \alpha_t(i) a_{ij} \sum_{d=1}^{D} P(d|j) b_j\left(O^{t'}_{t+1}\right) \beta_{t'}(j)}{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} \sum_{d=1}^{D} P(d|j) b_j\left(O^{t'}_{t+1}\right) \beta_{t'}(j)} \\
&= \frac{\sum_{t=1}^{T} \xi_{t,t'}(i,j)}{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_{t,t'}(i,j)}
\end{aligned} \tag{18}$$

4.4.3 Observation distributions

The re-estimation formula for segmental observation distributions is the expected number of times that observation $o_t{=}v_k$ occurred in state i, normalized by the expected

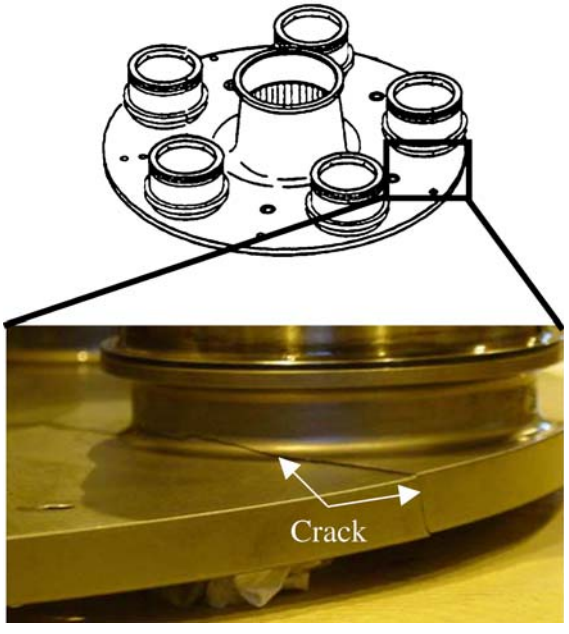


Fig. 4 Planetary carrier schematic and crack (Keller and Grabill [11])

Table 1 Accelerometers

No.	Name	Location	Axis
1	ACC1	Left accessory module	Vertical
2	ACC2	Right accessory module	Vertical
3	Input1	Left input module	Vertical
4	Input2	Right input module	Vertical
5	PortRing	Main module, left side	Radial
6	StbdRing	Main module, right side	Radial

number of times that any observation occurred in state i. Since $\alpha_t(i)$ accounts for the partial observation sequence $o_1 o_2 ... o_t$ and state i at t, while $\beta_t(i)$ accounts for the partial observation sequence $o_{t'} o_{t'+1} ... o_T$, given state i at t. The remainder of the observation sequence $o_t o_{t+1} ... o_{t'}$ given state i at t and state j at t' is accounted by $P(O_{t+1}^{t'} | t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j)$. Therefore, the re-estimation of segmental observation distributions can be calculated as follows:

$$\overline{b}_i(k) = \frac{\sum_{\substack{t=1 \\ s.t. O_t = v_k}}^{T} \alpha_t(i) P\left(O_{t+1}^{t'} | t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j\right) \beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i) P\left(O_{t+1}^{t'} | t=q_n, s_t=i, t'=q_{n+1}, s_{t'}=j\right) \beta_t(i)}$$

$$= \frac{\sum_{\substack{t=1 \\ s.t. O_t = v_k}}^{T} \alpha_t(i) \left[\frac{\phi_{t,t'}(i,j)}{\sum_{d=1}^{D} P(d=t'-t|i)} \right] \beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i) \left[\frac{\phi_{t,t'}(i,j)}{\sum_{d=1}^{D} P(d=t'-t|i)} \right] \beta_t(i)} \tag{19}$$

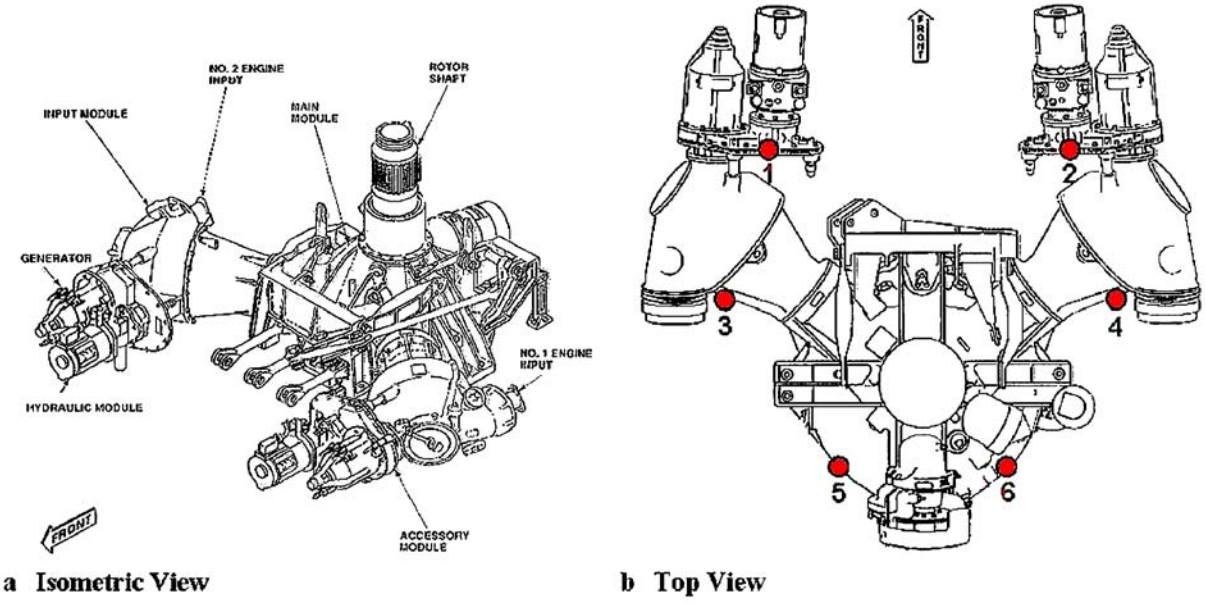
4.4.4 State duration probability distributions

The re-estimation problem is more difficult for HSMMs than for the standard HMM. One proposal to alleviate some of these problems is to use a parametric state duration density instead of the nonparametric duration density. A parametric model requires far less training and generalizes better results. In this study, Gaussian distribution is adopted. Gaussian distributions have many convenient properties, so random variates with unknown distributions are often assumed to be Gaussian. It is often a good approximation due to a surprising result known as the central limit theorem. This theorem states that the mean of any set of variates with any distribution having a finite mean and variance tends to the Gaussian distribution. In this paper, state duration probabilities are estimated directly from the training data.

The mean $\mu(i)$ and the variance $\sigma^2(i)$ of duration probability of health state i are determined by

$$\mu(i) = \frac{\sum_{q_{n-1}=1}^{T} \sum_{q_n=q_{n-1}}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left\{\frac{[(q_n - q_{n-1}) - \mu]^2}{2\sigma^2}\right\}} \cdot (q_n - q_{n-1})}{\sum_{q_{n-1}=1}^{T} \sum_{q_n=q_{n-1}}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left\{\frac{[(q_n - q_{n-1}) - \mu]^2}{2\sigma^2}\right\}}} \tag{20}$$

$$\sigma^2(i) = \frac{\sum_{q_{n-1}=1}^{T} \sum_{q_n=q_{n-1}}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left\{\frac{[(q_n - q_{n-1}) - \mu]^2}{2\sigma^2}\right\}} \cdot (q_n - q_{n-1})^2}{\sum_{q_{n-1}=1}^{T} \sum_{q_n=q_{n-1}}^{T} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left\{\frac{[(q_n - q_{n-1}) - \mu]^2}{2\sigma^2}\right\}}} - \mu^2(i) \tag{21}$$



a Isometric View **b Top View**

Fig. 5 UH-60A transmission and sensor locations [11]

4.4.5 Classification using HSMMs

For diagnostics, the goal is to develop trained HSMMs to recognize N different states of a component for a given failure mode. That is, the task is to develop a diagnostics model for classifying the health states of a component. Therefore, for the diagnosis of the fault, it is necessary that a separate HSMM be trained for all possible fault types in addition to the HSMM for normal conditions.

Given these N groups of observation sequences, N different HSMMs (i.e., $HSMM_1$, $HSMM_2$, ..., $HSMM_N$) are modeled for characterization of each group, which corresponds to a health state.

The classification of a failure mode given an observation sequence is done by presenting each of the N trained HSMMs with the same observation sequence. The HSMM that gives the highest log likelihood value represents the type of fault that is hidden in the observation sequence.

4.5 HSMM-based prognosis

The objective of prognostics is to predict the progression of a fault condition to system failure and estimate the remaining useful life (RUL) of the system. In the following, the HSMM-based prognosis procedure is described (Fig. 3).

Step 1 From the HSMM training procedure (i.e., parameter estimation), we can obtain the state transition probability of the HSMM.

Table 2 Diagnosis results of input1 and input2 based on HSMM

Transmission	Input1		Input2	
	$HSMM_1$	$HSMM_3$	$HSMM_1$	$HSMM_3$
1	26.4093	−Inf	22.6181	−63.6801
1	26.4093	−Inf	22.6137	−66.2003
1	20.5733	−102.098	23.9286	−242.4071
1	20.511	−172.443	23.9406	−199.6775
1	20.7164	−106.45	23.9319	−192.5113
1	20.5387	−149.703	22.3184	−201.9975
1	20.5909	−204.083	22.3606	−122.9994
1	20.375	−117.235	22.35	−213.83
2	−166.978	−477.403	−175.3629	−231.995
2	−162.986	−299.425	−124.4743	−158.9221
2	−109.667	−240.366	−240.5964	−415.6203
2	−93.6992	−254.437	−412.9129	−743.0538
2	−103.512	−174.024	−195.6991	−336.2356
2	−122.703	−157.782	−372.4606	−621.1884
3	−229.071	9.0944	−380.4721	9.9441
3	−283.546	8.7355	−354.0287	9.6233
3	−486.871	12.2049	−302.795	10.1132
3	−428.854	12.3552	−313.6242	10.4206
3	−402.364	12.3369	−310.159	11.0638
3	−369.06	12.2157	−317.1094	11.2319
3	−323.981	8.9392	−285.9819	10.4208
3	−369.158	9.1104	−316.8308	10.1274
3	−217.618	9.0602	−167.3965	10.3964
3	−310.124	8.5934	−61.3788	9.7423
3	−255.26	8.6324	−115.6322	10.3859
3	−403.99	12.3078	−95.8976	10.229
3	−311.591	12.3978	−140.0857	10.171
3	−359.93	12.4574	−98.0043	10.3317
3	−195.277	9.0026	−114.3106	9.9223
3	−130.057	8.7274	−122.7489	9.7185
4	−168.759	−66.3279	−250.8073	−297.594
4	−152.863	−49.481	−329.5202	−360.4773
4	−240.705	−13.6603	−119.11	−92.8467
4	−199.522	−15.8828	−161.3985	−108.9714
4	−240.705	−13.6603	−119.11	−92.8467
4	−199.522	−15.8828	−161.3985	−108.9714

Table 3 Diagnosis results of input1 and input2 based on HMM

Transmission	Input1		Input2	
	HMM_1	HMM_3	HMM_1	HMM_3
1	35.3502	−Inf	35.3502	−Inf
1	35.3502	−Inf	−208.107	−Inf
1	−Inf	−337.964	−267.6859	−Inf
1	−Inf	−Inf	35.3502	−Inf
1	−Inf	−333.26	−242.2206	−Inf
1	−Inf	−617.803	−198.0565	−Inf
1	−Inf	−Inf	−277.2974	−Inf
1	−Inf	−Inf	−216.0492	−Inf
2	−Inf	−Inf	−Inf	−Inf
2	−Inf	−Inf	−Inf	−Inf
2	−Inf	−Inf	−Inf	−Inf
2	−Inf	−Inf	−Inf	−Inf
2	−Inf	−701.641	−Inf	−Inf
2	−Inf	−619.024	−Inf	−Inf
3	−Inf	34.9448	−Inf	34.6556
3	−Inf	−78.9251	−Inf	34.6556
3	−Inf	−325.599	−Inf	34.5544
3	−Inf	−290.556	−Inf	34.5222
3	−Inf	−182.549	−Inf	34.2386
3	−Inf	−131.985	−Inf	27.2313
3	−Inf	−444.933	−Inf	−46.5141
3	−Inf	−401.866	−Inf	−64.0732
3	−Inf	−67.0239	−Inf	−265.577
3	−Inf	−129.268	−465.0761	−301.767
3	−Inf	−88.7317	−Inf	−316.175
3	−Inf	−205.153	−Inf	−319.472
3	−Inf	−206.716	−Inf	−312.348
3	−Inf	−166.377	−Inf	−363.797
3	−Inf	35.1458	−Inf	−510.351
3	−Inf	35.1458	−705.6241	−Inf
4	−Inf	−446.035	−Inf	−Inf
4	−Inf	339.772	−Inf	−Inf
4	−Inf	−271.334	−642.5215	−Inf
4	−Inf	−180.838	−Inf	−Inf
4	−Inf	−271.334	−642.5215	−Inf
4	−Inf	−180.838	−Inf	−Inf

Step 2 Through the HSMM parameter estimation, the duration mean and variance for each state can be estimated.

Step 3 By classification, identify the current health state of the system.

Step 4 The RUL of the system can be computed by following backward recursive equations (suppose that the system currently stays at health state i, RUL_i indicates the RUL starting from state i):

At state $N-1$:

$$RUL_{N-1} = a_{N-1,N-1}[D(h_{N-1}) + D(h_N)] + a_{N-1,N}[D(h_N)]$$

At state $N-2$:

$$RUL_{N-2} = a_{N-2,N-2}[D(h_{N-2}) + RUL_{N-1}] + a_{N-2,N-1}[RUL_{N-1}]$$

At state i:

$$RUL_i = a_{ii}[D(h_i) + RUL_{i+1}] + a_{i,I+1}[RUL_{i+1}] \quad (22)$$

5 Case studies

5.1 UH-60A Blackhawk main transmission planetary carrier fault

The H-60 main transmission employs a 5-planet epicyclic gear train. Recent inspections of two Army UH-60A Blackhawk main transmissions, initiated by indications of low or fluctuating oil pressure, revealed a crack in the planetary carriers. This is the first time this problem has been encountered in approximately 3.6 million UH-60A flight hours in the Army. The 10-inch crack in the first carrier, which was later cut apart for material inspection, is shown in Fig. 4. Since the planetary carrier is a flight critical part, failure could cause an accident resulting in loss of life and/or aircraft. This resulted in flight restrictions on a significant number of Army UH-60As. Manual inspection of all 1000 transmissions is not only costly in terms of labor, but also time prohibitive. A simple, cost-effective test capable of diagnosing this fault is highly desirable.

The U.S. Army Aviation Engineering Directorate, Aeromechanics Division conducted an investigation to determine if a fault in the planetary carrier of an H-60 transmission could be successfully detected via vibration monitoring. Experimental measurements of UH-60A transmissions with both faulted and unfaulted planetary carriers were taken in a test cell and on-aircraft.

Test cell measurements The Helicopter Transmission Test Facility (HTTF) at the Naval Air Station Patuxent River,

Table 4 Summary of diagnostic results using both HSMMs and HMMs

Measurements	Accelerometers and their locations	HSMM classification rate(%)	HMM classification rate(%)
Test cell measurements	ACC1	100	80
	ACC2	100	70
	Input1	100	60
	Input2	100	73.33
	PortRing	100	70
	StbdRing	100	60
On-aircraft measurements	Input1	100	45.83
	Input2	100	95.83
	PortRing	100	50
	StbdRing	100	29.17

Table 5 Transition probability between four health states

States	Baseline	Contamination1	Contamination2	Contamination3
Baseline	0.8913	0.0454	0.0000	0.0633
Contamination1	0.0000	0.6399	0.3599	0.0003
Contamination2	0.0000	0.0000	0.9167	0.0833
Contamination3	0.0000	0.0000	0.0000	1.0000

MD was utilized for all test cell measurements. The HTTF is a unique test facility that uses actual aircraft engines to provide power to all the aircraft drive systems except the rotors and is a significant improvement over single component test rigs. An H-60 test transmission was instrumented with several sets of accelerometers. The accelerometers and their locations are included in Table 1 and shown in Fig. 5.

The vibration data was acquired using the US Army's Vibration Management Enhancement Program (VMEP) system. Measurement setups were created to simultaneously acquire and synchronize the six input accelerometers to the carrier rotational frequency. The VMEP system was programmed to calculate and save the raw time domain data, the time synchronous waveform, and the condition indicators. Time synchronous vibration data were measured for each accelerometer at torque settings ranging from 20 to 100%. Both faulted and unfaulted planetary carriers were tested. The faulted planetary carrier, the second cracked UH-60A carrier discovered during a field inspection, had a 314-inch crack when installed in the test transmission. Because of the need for destructive material testing on the cracked carrier, the amount of run time was limited. Thus, the transmission was not run for an extended period. Only "snapshots" of data at each torque setting were measured.

On-aircraft measurements Similar to the test cell measurements, several UH-60A transmissions were instrumented. However, for the on-aircraft tests only accelerometers 3 through 6 were used. Time synchronous vibration data were measured for each accelerometer using the same data acquisition equipment that was used in the test cell. The same faulted carrier and main gearbox used in the tests at the HTTF was installed in a UH-60A at the Corpus Christi Army Depot (CCAD) and used as a test aircraft. Three different UH-60As from the Birmingham, Alabama National Guard (BNG) with unfaulted carriers were selected as test aircraft to establish how vibration levels change across aircraft. Because of safety issues with the cracked carrier only ground runs were completed. Single and dual engine data were taken at torque settings of 20%Q and 30%Q. For each test state, the aircraft was stabilized for 5 min before vibration measurements were taken. For all tests, winds were less than 10 knots with the nose of the aircraft pointed into the wind.

These vibration signals were processed using wavelet packet with db10 wavelet and five decomposition levels. The wavelet coefficients obtained by the wavelet packet decomposition were used as the inputs to the HMMs and HSMMs. In this test, we wanted to see how the HSMMs could classify the health conditions of the transmissions in comparison with the HMMs.

5.2 Diagnostic results

The experiment and collection procedure of experimental data can be found in [11]. Provided in Table 2 and Table 3 are the diagnostic results of on-aircraft data using HSMMs and HMMs, respectively. Here, only the results for accelerometers input 1 and input 2 are provided. The values in both tables are the log likelihood values computed by either HSMMs or HMMs. When the log likelihood value approaches negative infinitive, it is represented as "−Inf" in both tables. As the on-aircraft data were generated from four different transmissions, four different HSMMs ($HSMM_1$, $HSMM_2$, $HSMM_3$, $HSMM_4$) and four different HMMs (HMM_1, HMM_2, HMM_3, HMM_4) were obtained. However, only diagnostic results for transmissions 1 and 3 are provided.

In Table 2, $HSMM_1$ and $HSMM_3$ represent the HSMMs trained using the data from transmission 1 and transmission 3, respectively. From Table 2, we can see that if we present $HSMM_1$ with data obtained from each transmission, $HSMM_1$ gives larger log likelihood values for the data from transmission 1 than the data from other transmissions. This indicates that $HSMM_1$ is able to differentiate the signals from faulted transmission (i.e., transmission 1 in this case) and unfaulted transmissions. The same interpretations apply to $HSMM_3$ in Table 2.

In Table 3, HMM_1 and HMM_3 represent the HMMs trained using the data from transmission 1 and transmission 3, respectively. From Table 3, we can see that if we present HMM_1 with data obtained from each transmission, HMM_1 cannot give larger log likelihood values for all the data from transmission 1 than the data from other transmissions. This indicates that HMM_1 is unable to differentiate the

Table 6 Mean and variance of duration time in four health states

States	Baseline	Contamination1	Contamination2	Contamination3
Mean of duration	10.4549	9.7923	11.3375	10.4793
Variance of duration	1.9388	0.9792	1.2415	0.1880

signals from faulted transmission (i.e., transmission 1 in this case) and unfaulted transmissions. The same interpretations apply to HMM_3 in Table 3.

The diagnostic results using both HSMMs and HMMs are summarized in Table 4. From Table 4, we can see that the HSMMs have much better performance in diagnosing the fault of the transmission in comparison with the HMMs.

5.3 Prognosis for pumps

The prognosis was tested using data from a real hydraulic pump health monitoring application case study [7]. In this case study, three pumps (pump 6, pump 24, and pump 82) were run under four different testing conditions: baseline (normal state), contamination 1 (20 mg of dust injected into the oil reservoir), contamination 2 (40 mg of dust injected into the oil reservoir), and contamination 3 (60 mg of dust injected into the oil reservoir). The contamination stages in this hydraulic pump wear test case study correspond to different stages of flow loss in the pumps. As flow rate of a pump clearly indicates the heath state of a pump, therefore, the contamination stages corresponding to different degrees of flow loss in a pump were defined as the health states of the pump in the pump wear test. During the test run, under each condition, vibration signals were collected. These signals were processed using wavelet packet with db10 wavelet and five decomposition levels. The wavelet coefficients obtained by the wavelet packet decomposition were used as the inputs to the HMMs and HSMMs.

The lifetime training data from pump 6, pump 24, and pump 82 were used for prognostics study. By training, an HSMM with four health states can be obtained. And, the mean and variance of the duration time in each state are also available through the training process. The results are given in Tables 5 and 6. Based on the above information, the mean value of the RUL of a pump can be calculated as follows (in terms of Eq. 22 and suppose that the component currently stays at state "Contamination1"):

$$Mean\_RUL_{c1} = 33.4004$$

Similarly, the variance of the remaining useful life of a pump can be obtained as follows:

$$Variance\_RUL_{cl} = 1.9528$$

That is, if the component is currently at state "Contamination1," then its expected remaining useful life is 33.4004 time units with a confidence interval of 1.9528 time units.

6 Conclusions

In this paper, we have presented an integrated methodology for equipment health diagnosis and prognosis. The presented methodology is developed based on hidden semi-Markov models (HSMMs). In this proposed HSMM-based diagnostics and prognostics methodology, health states of equipment are modeled by state transition probability matrix and observation probability. The duration of each health state is described by the state duration probability.

To facilitate the computation of the HSMMs, new forward and backward variables are defined. And correspondingly, the re-estimation formulae based on new variables are used. A modified forward-backward algorithm is developed to estimate the parameters of the HSMMs. By incorporating the explicit temporal structure, the diagnostic and prognostic power of the HSMMs is improved.

The evaluation of the HSMM-based methodology was carried out by two real-world application case studies: fault diagnosis of UH-60A Blackhawk main transmission planetary carriers and prognosis of a real hydraulic pump health monitoring application. For test cell measurements, the correct recognition rate is increased by 45% compared with the HMMs; for on-aircraft measurements, the correct recognition rate is increased by 81% compared with HMMs. These results show the effectiveness of the HSMM-based diagnostic approach. On the other hand, it becomes possible for us to use the same HSMMs to predict the remaining useful life of the equipment.

References

1. Atlas L, Ostendorf M, Bernard GD (2000) Hidden Markov models for monitoring machining tool-wear. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing–Proceedings, vol 6, pp 3887–3890
2. Baruah P, Chinnam RB (2003) HMMs for diagnostics and prognostics in machining processes. Proc. of the 57th Society for Machine Failure Prevention Technology Conference, Virginia Beach, VA, 14–18 April
3. Begg CD, Merdes T, Byington C, Maynard K (1999) Dynamic modeling for mechanical diagnostics and prognostics. Maintenance and Reliability Conference (MARCON99)
4. Bunks C, Mccarthy D, Tarik A (2000) Condition based maintenance of machines using hidden Markov models. Mech Syst Signal Process 14(4):597–612
5. Chen MY, Kundu A, Zhou J (1994) Off-line handwritten work recognition using a hidden Markov model type stochastic network. IEEE Trans Pattern Anal Mach Intell 16(5):481–496
6. Chen MY, Kundu A, Srihari SN (1995) Variable duration hidden Markov model and morphological segmentation for handwritten word recognition. IEEE Trans Image Process 4 (12):1675–1688
7. Dong M, He D (2004) Hidden semi-Markov models for machinery health diagnosis and prognosis. Trans NAMRI/SME XXXII:199–206
8. Ertunc HM, Loparo KA (2001) A decision fusion algorithm for tool wear condition monitoring in drilling. Int J Mach Tools Manuf 41:1347–1362

9. Ertunc HM, Loparo KA, Ocak H (2001) Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs). Int J Mach Tools Manuf 41:1363–1384
10. Kannan A, Ostendorf M (1993) Comparison of trajectory and mixture modeling in segment-based word recognition. Proceedings-ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, vol 2, Speech Processing, pp 327–330
11. Keller JA, Grabill P (2003) Vibration monitoring of UH-60A main transmission planetary carrier fault. American Helicopter Society 59thAnnual Forum, Phoenix, AZ, 6–8 May 2003
12. Ljolie A, Levinson SE (1991) Development of an acoustic-phonetic hidden Markov model for continuous speech recognition. IEEE Trans Signal Process 39(1):29–39
13. Ostendorf M (1989) Stochastic segment model for phoneme-based continuous speech recognition. IEEE Trans Acoustics Speech Signal Process 37(12):1857–1869
14. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77 (2):257–286
15. Roemer MJ, Kacprzynski GJ (2000) Advanced diagnostics and prognostics for gas turbine engine risk assessment. Proceedings of the 2000 IEEE Aerospace Conference, Big Sky, MT, 18–25 March
16. Roemer MJ, Nwadiogbu EO, Bloor G (2001) Development of diagnostic and prognostic technologies for aerospace health management applications. Proceedings of the 2001 IEEE Aerospace Conference, Big Sky, MT, 10–17 March
17. Wang L, Mehrabi MG , Kannatey-Asibu E Jr (2002) Hidden Markov model-based tool wear monitoring in machining. ASME J Manuf Sci Eng 124:651–658
18. Russell MJ, Moore RK (1985) Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. Proc. of ICASSP '85, Tampa, FL, pp 5–8