Fault Prognosis of Key Components in HVAC Air Handling Systems at Component and System Levels

Ying Yan, *Student Member*, *IEEE*, Peter B. Luh, *Life Fellow*, *IEEE*, Krishna R. Pattipati, *Life Fellow*, *IEEE*

***Abstract*— Fault prognosis of air handling systems, which are key sub-systems of Heating, Ventilation and Air Conditioning systems, allows system operators to know the Remaining Useful Life (RUL), thus preventing unexpected breakdowns and reducing operational and maintenance costs. In this paper, a new hidden Semi-Markov model-based method is developed. In the method, only relevant state transition points are selected and estimated, leading to computational efficiency. Physics-based models are used in a novel way to provide "mapping matrices" relating component capacities to fault severities, capturing impacts of multiple failure modes. Experimental results show that our method can effectively estimate RUL of components and systems.**

***Note to practitioners* —Air handling systems within HVAC condition the air to satisfy human thermal comfort and air quality requirements. The fault prognosis of these systems is critical since it allows system operators to know the Remaining Useful Life (RUL), thus preventing unexpected breakdowns and reducing operational and maintenance costs. In this paper, a new hidden Semi-Markov model-based method is developed to estimate RUL of an air handling system and its components. Experimental results show that our method can predict RUL of components and systems with high accuracy.**

***Index Terms*— Air handling system, fault prognosis, remaining useful life, hidden semi-Markov model**

I. Introduction

HEATING, Ventilation, and Air-Condition (HVAC) systems constitute 57% of the energy used in the U.S. commercial and residential buildings [1]. Air-handling systems are key subsystems of HVAC systems that condition the air to satisfy the human thermal comfort and air quality requirements. Variable Air Volume (VAV) air handling systems are the most common ones used today, and are comprised of components including 1) dampers; 2) filters; 3) cooling/heating coils; 4) fans; and 5) ducts, as shown in Fig. 1.

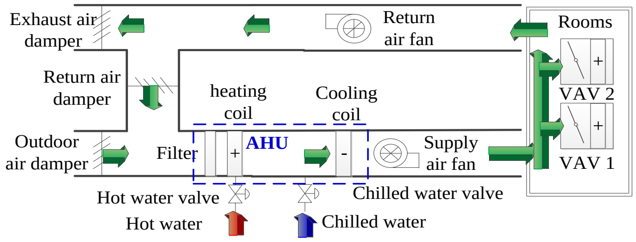

Fig. 1. Structure of a specific air handling system considered

The recirculation air damper and the outdoor air damper mix the air in the desired proportion. The mixed air then passes through coils to condition the air via heat exchange. The supply fan delivers air to VAV boxes. The return fan delivers [1]air to the exhaust air and the return air damper. Components of the air handling systems typically have long lives [2].

In air handling systems, operational capacities of components and systems determine how well human comfort and air quality requirements are satisfied. Degradation in component capacities thus captures effects of their faults. Sudden (abrupt) and gradual faults constitute two major categories of faults in components. Sudden faults occur unexpectedly and are unpredictable. Gradual faults, e.g., aging of equipment, cause a slow degradation in component and system capacities, and are predictable. Because of gradual faults, a component or a system may no longer perform its intended function over time. The remaining time that a component or a system is able to function in accordance with its intended purpose is termed its Remaining Useful Life (RUL). Fault prognosis allows operators to know the RUL of components and systems, thereby preventing unexpected breakdowns and reducing operational and maintenance costs. Prognosis at the component-level helps in determining when components need repair or replacement; prognosis at the system-level provides a global view of the system health. However, prognosis is challenging because 1) estimating conditions of components and systems with low false identification rates may require high computational effort, leading to a slow inference; and 2) models capturing impacts of multiple failure modes may be too complex to be established.

To keep a component or a system running normally, a fault diagnosis method identifies current failure modes and their severities. If a fault is isolated, maintenance crews will repair or replace the failed component or system. Otherwise, the fault prognosis method estimates RULs based on current conditions. Thus, it is important to estimate the current conditions rapidly. Hidden Markov Models (HMM) and Hidden Semi-Markov Models (HSMMs) are usually used to estimate current conditions. In these methods, their model parameters are estimated first, and are re-estimated at intervals to adapt to changing environments. Given parameters, their states, i.e., current conditions, are estimated. Compared to HMMs, HSMMs have explicit time-duration distribution for each state, thus can capture more general time-evolution of degradations than HMMs can. However, the Viterbi algorithm used to estimate parameters and states of HSMMs needs to

The research was supported in part by the National Science Foundation under Grant CCF-1331850. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Y. Yan is with the School of Automation, Nanjing University of Information Science and Technology. P. B. Luh and K. R. Pattipati are with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269, USA.

calculate probabilities of every possible time-durations in each state and at each time epoch iteratively, thus is time-consuming (e.g., it requires 5.6 hours in our problem). A fault is missed if it occurs when parameters are being estimated.

This paper aims to address the problem of RUL predictions at both the component and the system levels for air handling systems by employing HSMMs. Since cooling coils and supply fans are used to condition and circulate air to rooms and are the key components in air handling systems, they constitute as examples to illustrate our method. However, our method easily extends to other components as well. Based on accepted practice in HVAC systems [3], tube fouling and dust on fins in cooling coils, and a decrease in fan efficiency comprise the failure modes to illustrate our method. Since tube fouling mainly depends on the quality of chilled water and dust on fins mainly relies on air quality, they are considered independent of each other. This paper focuses on the following two aspects as shown in Fig. 2. The first is a computationally efficient statistical method to estimate the parameters and states of HSMMs with low false identification rates. The second is a statistical method to estimate the RUL of a component or a system by combining the influence of related failure modes via mapping matrices.

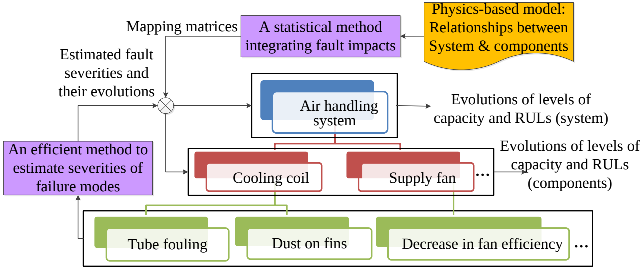
Mapping matrices
A statistical method integrating fault impacts
Physics-based model: Relationships between System & components
Estimated fault severities and their evolutions
Air handling system
Evolutions of levels of capacity and RULs (system)
An efficient method to estimate severities of failure modes
Cooling coil
Supply fan
...
Evolutions of levels of capacity and RULs (components)
Tube fouling
Dust on fins
Decrease in fan efficiency
...

Fig. 2. The framework of our fault prognosis method

The remaining sections of the paper are as follows. Section 2 provides a brief review of the fault prognosis methods in the literature. In Section 3, an accurate and computationally efficient HSMM-based fault prognosis method is developed to estimate the RUL of components with a single failure mode. In Section 4, the HSMM method is extended to estimate the RUL of components with multiple failure modes. In Section 5, our fault prognosis method is tested using data from a small simulated building. Experimental results show that our method performs well in estimating the RULs at both the component and system levels.

II. Literature Review

There is a paucity of literature on the fault prognosis of HVAC air handling systems and their components. Consequently, methods developed for other components (e.g., engines) are reviewed in subsection II-A. Combining component-level results to system-level prognosis involves additional layers of difficulty, as reviewed in subsection II-B.

A. Fault Prognosis Methods for Components

Physics-based, black box and statistical models constitute three categories of prognosis methods. First principles-based mathematical models of component degradations form the basis for the prognosis of physics-based methods. In [4-6], extended Kalman filter and particle filter-based methods constitute the basis for estimating the parameters of time-dependent degradation models.

As typical black-box models, static neural networks (NNs), can approximate arbitrary nonlinear relationships, but they do not capture temporal evolutions, and thus are not suitable for fault prognosis. Unlike static NNs, recurrent NNs represent the output at time *t* as a function of previous output and external inputs at the current and previous times, and are suitable for fault prognosis [7]. In [8], a neuro-fuzzy network-based fault prognosis scheme predicted the spur gear condition one-step ahead. The fuzzy inference structure was established based on domain knowledge, and the concomitant fuzzy membership functions were trained by NNs. In [9], fault prognosis in traction motors used in wind turbines employs torque prediction via a least-squares support vector regression method as a prognostic indicator.

Statistical models, e.g., dynamic Bayesian networks and HMMs, represent time evolution statistically. In these models, the time-to-failure is assumed to follow a certain distribution, e.g., geometric distribution. For instance, a dynamic Bayesian network model was used to estimate the RUL of a computer numerical control tool machine [10]. In [11], to estimate RUL of a component under degradation or shock damage, the degradation time was modeled as Gaussian, and the time between two consecutive shock damages was modeled as a Poisson distribution. HMM-based methods are used to estimate the RUL of components, e.g., bearings [12]. However, in HMMs, the time-duration distribution of each state is assumed to be geometric, and this assumption may not be realistic in practice. HSMM is an extension of HMM by allowing the underlying process to be a semi-Markov chain with time-duration distributions for each state. Compared to HMMs, HSMMs have explicit time-duration distributions and thus can capture the state evolutions more accurately. In [13], a segmental HSMM-based method was used to predict the RUL of pumps. However, unlike HMMs, estimating HSMM parameters and states needs to calculate probabilities of every possible time-duration for each state at each time epoch, leading to high computational cost. Table I summarizes the fault prognosis methods for components.

Table 1

Comparison of Fault Prognosis Methods

Approaches	Components/ systems	Advantages	Disadvantages
Physics-based models [4-6]	Pump, gearbox, bearings, etc.	More precise than others,	Physical knowledge may not be available
Black-box models [7-9]	Gearbox, bearings, engine, etc.	Applicable without physical knowledge	Lack of transparency and robustness
Statistical model-based [10-13]	Pump, bearings, etc.	Robust and easy to establish	May be time-consuming

B. Fault Prognosis Methods for Systems

Some papers carried out fault prognosis of systems by identifying the rate of change in damage indicators, e.g., in planes [14]. Some other papers considered fault prognosis by estimating RULs. Typically, the system prognosis is assumed to depend upon a few critical components. Tracking the remaining life of these critical components provides a measure of the remaining life of the entire system [15]. For instance, the prognosis of a suspension system depends on crack growth on the suspension spring which is a critical component in the suspension [15]. Some other methods consider the combined impacts of all components on the system. For instance, in [16], a hierarchical architecture was used to analyze the effects of system-level parameters on component faults in an aero propulsion system of turbofans. In [17], system-level performance was calculated based on health factors of components to predict RULs of an aircraft's air conditioning system. These methods require complex models to capture relationships between components and systems, thus are hard to use.

III. Estimating RULs of Components under One Failure Mode

In subsection III-A, estimating the RUL of a component with one failure mode is considered first. An HSMM is established to capture the transitions among fault severities, with their parameters estimated via Gibbs sampling. In subsection III-B, an efficient statistical method with low false identification rates and low computational effort is developed to infer the states of HSMMs. In subsection III-C, an improved backward recursion method is derived to estimate the RUL based on the estimated parameters and states.

A. HSMMs Capturing Transitions among Fault Severities

To establish an HSMM for the severities of a failure mode, impacts of the fault are analyzed first. Based on the analysis, appropriate observations are selected for the HSMM. Then, parameters of the HSMM are estimated. To illustrate the process, a single failure mode of a supply fan is used as an illustrative example. Supply fans deliver air to rooms. The fan capacity is a function of the supply fan efficiency e_{sf}. A decrease in fan efficiency e_{sf} results in the fan consuming more electricity to maintain a specified $\dot{m}_{a,\sup}$. Based on this concept, a physics-based empirical model of a fan is derived as follows [18]:

$$\frac{Q_{sf}\cdot\rho_{air}\cdot e_{sf}}{\dot{m}_{\sup,des}\cdot\Delta P_{sf}} = c_1 + c_2 f_{sf} + c_3 f_{sf}^2 + c_4 f_{sf}^3 + c_5 f_{sf}^4 \text{, with} \quad (1)$$

$$f_{sf} = \dot{m}_{a,\sup} \big/ \dot{m}_{\sup,des}\,. \quad (2)$$

where Q_{sf} is the power of the supply fan; ρ_{air} is the air density; ΔP_{sf} is the pressure rise through the supply fan; $\dot{m}_{\sup,des}$ is the design value of $\dot{m}_{a,\sup}$; and c_1, c_2, c_3, c_4, and c_5 are coefficients. Since the decrease in e_{sf} reflects performance degradation, it is considered as the degradation state of the supply fan, and a fault indicator. As presented in [13], the supply fan is assumed to have four states S_{sf} = 0, 1, 2 and 3, including a normal condition (S_{sf} = 0) and three severities of the gradual fault (S_{sf} = 1, 2 and 3), where Fault Severity 3 is considered as the unaccepted level, as shown in Fig. 3. An HSMM is characterized by four parameters: 1) the initial probability vector π; 2) the state transition matrix P; 3) the observation symbol probability distribution $B = \{b_i(O(k))\}$, where $b_i(O(k))$ is the probability that the observation O at time k belongs to State i, governing distributions of observations; and 4) a parameter set λ containing parameters of time-duration distributions in each state [13].

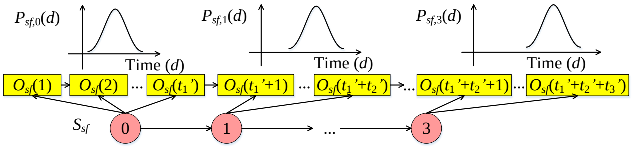

Fig. 3. HSMM representing state evolutions of a supply fan

To select appropriate observations for HSMMs, as discussed in [19], three kinds of variables are considered, including 1) fault-related sensor readings; 2) errors between some sensor readings and set-points; and 3) residuals between some sensor readings and their predictions obtained from models. For instance, for the supply fan, the supply air mass flow rate $\dot{m}_{a,\sup}$ may decrease because of a decrease in e_{sf}. Since the fan consumes more electricity under the faulty condition, temperature rise ΔT_{sf} of the air through the fan may increase. The residual $R_{phy,sf}$ between the left-side and the right-side of (1) approximates to zero under normal conditions, but increases with a decrease in e_{sf}. Thus, the three fault-related variables compose the fault indicator matrix as:

$$X_{sf} = \begin{bmatrix} \dot{m}_{a,\sup}^1 & \Delta T_{sf}^1 & R_{phy,sf}^1 \\ . & . & . \\ \dot{m}_{a,\sup}^K & \Delta T_{sf}^K & R_{phy,sf}^K \end{bmatrix}, \quad (3)$$

where K is the length of the observation sequence. These variables may be related and contain redundant information. To remove redundancy, principal component analysis is used. For X_{sf}, the first component captures 98.052% of variation in the data and is used as the observation O_{sf}.

To estimate parameters of HSMMs, the expectation-maximization algorithm (EM) and Gibbs sampling are usually used. Unlike the EM, Gibbs sampling is a Markov chain Monte Carlo algorithm. In the method, parameters are sampled randomly based on their posterior distributions conditioned on other parameters (like in a Gauss-Seidel iteration) until the iterations converge. It ensures that a Markov chain converges to a stationary distribution that is the distribution of HSMM parameters. Since Gibbs sampling avoids local minima, it is used to estimate parameters of HSMMs in our problem [20]. Both HMMs and HSMMs have initial probability vectors, state transition matrices, and emission probabilities, and thus posterior distributions used for HMMs are used for HSMMs [19]. Posterior distributions of the three parameters are as (9)-(16) in [19]. Additionally, HSMMs have extra parameters of time-duration distributions. As presented in (7) of [21], the posterior distributions of the time-duration parameters are

$$P(d_i) \propto e^{-\lambda_i}\cdot\left(\lambda_i^{d_i} \big/ d_i!\right), \quad (4)$$

where d_i is the time-duration of State i; λ_i is the parameter of the Poisson distribution. The prior distribution of λ_i follows a gamma distribution as,

$$\lambda_i \sim \Gamma(\alpha_i,\beta_i) = \left(\beta_i^{\alpha_i}\lambda_i^{\alpha_i-1}e^{-\beta_i\lambda_i}\right)\Big/\Gamma(\alpha_i), \quad (5)$$

where α_i is the shape parameter, and β_i is the rate parameter of State i. The posterior distribution of λ_i also follows a gamma distribution as [21]

$$\lambda_i \mid \alpha_i, \beta_i \sim \Gamma(\alpha_i + l_i, \beta_i + m_i), \tag{6}$$

where m_i is the number of segments in State i; $l_i = \sum_{t=1}^{T} \delta_{S(t)}$, and $\delta_{S(t)}$ is the Kronecker delta function

$$\delta_{S(t)=i} = \begin{cases} 1, & s(t) = i \\ 0, & otherwise \end{cases}; \tag{7}$$

B. An Efficient Method to Estimate States of HSMMs with Low Computational Effort

As discussed before, to make a component perform its intended function, it is important to quickly estimate states of HSMM to know whether a fault occurs, and when a fault occurs in the future. The Viterbi algorithm is usually used to estimate HSMM parameters and states [22]. In this algorithm, at time t, the probability of observations corresponding to a state segment of staying in State i from time t-d+1 to time t needs to be calculated. This probability depends on 1) the probability of starting a time-duration in State i at time t-d+1; 2) the probability of staying in State i for a time-duration d; and 3) the probability that observations from time t-d+1 to time t belong to State i. The maximum value of this probability is calculated as:

$$\delta_i(t) = \max_d \left(\max_j \delta_j(t-d) a_{ji} \right) p_i(d) \prod_{s=t-d+1}^{t} b_i(O(s)), \tag{8}$$

where $p_i(d)$ is the probability that the time-duration of State i is d. As shown in (8), to obtain $\delta_i(t)$, probabilities of each state and each possible time duration should be calculated at each time. Since estimating states of HMMs does not need to calculate probabilities for each possible time-duration, estimating states of HSMMs is much more time consuming than that of HMMs. In our problem, the Viterbi algorithm runs iteratively with Gibbs sampling to estimate HSMM parameters, and the computation time is 5.6 hours. To adapt to changing environments, the parameters need to be re-estimated at intervals. A fault is said to be missed if it occurs when parameters are being estimated.

Components of air handling systems have long lives, thus they usually stay in the normal condition for a relatively long time. Therefore, state transitions are low probability events. Consequently, the number of state transition points is much smaller than that of non-transition ones. Since the non-transition points do not transit to other states, their estimates are the same as their previous transition point. If the transition points are estimated accurately, it is not necessary to estimate the states of non-transition points, leading to a substantial decrease in the computational effort. Based on this concept, a new statistical method is developed to estimate the states of HSMMs in an efficient manner as shown in Fig. 4. This method selects points that are more likely to transit to other states as potential transition points (black circles). L_1 points before and L_2 points after each selected point are included as potential state transition points (green triangles) to cover actual transition points. The rest of the points are considered as non-transition points (blue pentagons). Since only potential transition points need to be estimated and non-transition points are the same as their previous transition points, the computational requirements are reduced. In our method, actual state transition points may not be 100% covered by potential ones. If certain actual points are not covered, the estimation accuracy decreases. In our experiments, by selecting appropriate L_1 and L_2, all actual points are covered.

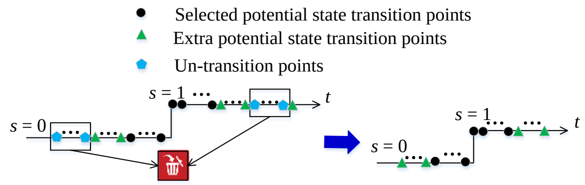
● Selected potential state transition points
▲ Extra potential state transition points
⬟ Un-transition points
s = 1 …
t
s = 0
s = 1
t
s = 0

Fig. 4. Selecting potential state transition points to cover all actual ones

The key issue here is to find state transition points in an efficient way. To determine potential transition points, it is important to calculate the transition probability and the non-transition probability for each point. For instance, if the current state is i at time t, the probability of transiting from State i to another state is

$$\upsilon_{ij}(t)=\Pr(O(1{:}t), S(1)=s_1,\ldots, S(t)=i, S(t+1)=j|_{j\neq i}), \tag{9}$$

and the probability of staying in State i is

$$\upsilon_{ii}(t)=\Pr(O(1{:}t), S(1)=s_1, \ldots, S(t)=i, S(t+1)=i). \tag{10}$$

As long as $\upsilon_{ij}(t) > \upsilon_{ii}(t)$ holds for at least one State j, the transition probability is deemed to be larger than the non-transition probability. Thus, let

$$\upsilon_i(t) = \max_j \left(\upsilon_{ij}(t) \right), \tag{11}$$

and the point at time t is more likely to transit to other states if

$$\upsilon_i(t) > \upsilon_{ii}(t). \tag{12}$$

The point satisfying (14) is considered as a potential state transition point. The probabilities $\upsilon_{ij}(t)$ and $\upsilon_{ii}(t)$ can be calculated by using the forward variable γ [22]. However, to do this, probabilities for each state and each possible time duration need to be calculated at each time epoch, leading to high computational requirements. To address this issue, the forward variable not considering probabilities of different possible time durations is used to approximate γ. For instance, the current state is j at time t, and thus the forward variable not considering time-durations is

$$\alpha_i(t) = \Pr\left(O(1:t), S(t) = i\right) \tag{13}$$

$$= \left[\sum_{j=0}^{N-1} \alpha_j(t-1) a_{ji} \right] \cdot b_i(O(t)), \tag{14}$$

Given (9), (10) and (13), it follows

$$\alpha_i(t) \approx \sum_{j \neq i} \upsilon_{ij}(t) + \upsilon_{ii}(t). \tag{15}$$

Then, it is easy to obtain

$$\sum_{j \neq i} \upsilon_{ij}(t) = \left[\sum_{j \neq i} \alpha_j(t-1) a_{ji} \right] \cdot b_i(O(t)), \text{ and} \tag{16}$$

$$\upsilon_{ii}(t) = \alpha_i(t-1) a_{ii} \cdot b_i(O(t)). \tag{17}$$

Based on (16) and (17), $\upsilon_{ij}(t+1)$ and $\upsilon_{ii}(t+1)$ are calculated. By checking (12), potential state transition points are determined. The computational complexity of this standard Viterbi algorithm is $O((ND^2+N^2)K)$, where N is the number of states; K is the length of the original state sequence; and D is the maximum duration allowed for any state. By using our method, the length of the state sequence is reduced from K to C, where C is the length of the shorter sequence. Thus, the computational complexity is reduced from $O((ND^2+N^2)K)$ to $O((ND^2+N^2)C)$. Since our paper focuses on fault prognosis, the sensitivity analysis of HSMMs is not considered.

C. An improved Backward Recursive Algorithm to Estimate RULs of Components under One Failure Mode

As mentioned in subsection III-A, each failure mode is assumed to have three fault severities. Fault Severity 3 is considered as the unacceptable level, and thus the time to Fault Severity 3 is regarded as the RUL. The state of the HSMM established in subsection III-A is the fault severity. Based on the estimated fault severity and parameters of the HSMM, a backward recursion was developed to estimate RULs [12]. This method assumes that the component will stay in the current state or transits from Severity *i* to Severity *i*+1. However, in practice, the component may transit to an even worse fault severity, e.g., from Severity *i* to Severity *i*+2. In our improved method, these ignored situations are covered as shown in Fig. 5.

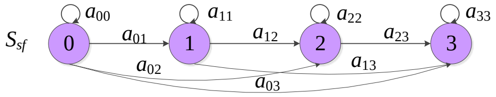

Fig. 5. Transitions among severities of decrease in the supply fan efficiency

In the figure, the supply fan could transit from State *i* to State *j*, where $i \leqslant j \leqslant 3$ since the upper bound of fault severities is 3 in our problem. Thus, RULs are calculated as follows:

$$\text{In State 3: } RUL_3 = 0, \tag{18}$$

$$\text{In State 2: } RUL_2 = a_{22} \cdot (D_2 - t_{stay,2}), \tag{19}$$

$$\text{In State 1: } RUL_1 = a_{11} \cdot (D_1 - t_{stay,1} + RUL_2) + a_{12} \cdot RUL_2, \tag{20}$$

$$\text{In State 0: } RUL_0 = a_{00} \cdot (D_0 - t_{stay,0} + RUL_1) + a_{01} \cdot RUL_1 + a_{02} \cdot RUL_2, \tag{21}$$

where RUL_i is the RUL of the supply fan in State *i*; the variable $t_{stay,i}$ is the length of time so far in State *i*; and D_i is the time duration in State *i*. As shown in (18), the RUL is zero for a component in State 3 (Fault Severity 3), since State 3 is deemed unacceptable. If the component is in State 2, it has two possible evolutions, including staying in State 2 for D_2-$t_{stay,2}$ time units with probability a_{22} or transiting to State 3 with probability a_{23} immediately. In the first case, the *RUL* is D_2-$t_{stay,2}$; in the second case, the *RUL* is 0. Thus, the summation is $a_{22} \cdot D_2$ as shown in (19). RULs of the supply fan in other states are computed similarly.

IV. Estimating RULs of Components and Systems under Multiple Failure Modes

In this section, estimating RULs of components or systems under multiple failure modes is considered. In subsection IV-A, mapping matrices are extracted from physics-based models to capture discrete relationships between states of a component and states of failure modes. In subsection IV-B, a statistical method is developed to estimate RULs of a component while considering the impacts of multiple failure modes, and the method is then extended to that of an air handling system.

A. Extract Mapping Matrices from a Physics-based Model

As discussed before, two failure modes of a cooling coil are considered: tube fouling and dust on fins. To estimate states (i.e., both failure modes and fault severities) of the cooling coil, HSMMs should capture the impacts of these two failure modes. By using the method presented in subsection III-A, two kinds of HSMMs are established. One kind is used to estimate failure modes with states $S_{falmod,cc}$ being combinations of failure modes, e.g., $S_{falmod,cc}$ = 1 representing (f_{tube}, f_{fin}) = (0, 1) which means that tube fouling does not occur and dust on fins occurs. The other kind is used to estimate severities of the two failure modes. To estimate RULs of the cooling coil, an HSMM describing state evolution of the cooling coil is required. States S_{cc} of the cooling coil are levels of capacity. Since the cooling capacity depends on severities of the two failure modes, S_{cc} relies on joint states S_{js} of the two failure modes. With each failure mode having four states, there are N_{js} = 4^2=16 joint states, e.g., S_{js} = 3 is equivalent to (S_{tube} = 0, S_{fin} = 3). With HSMMs of the two failure modes established, they are directly combined to create the HSMM for the cooling coil. To achieve the above, a discrete model describing the discrete relationship between HSMM states of failure modes and those of the cooling coil is required but is not available. To address this issue, a cooling coil's physics-based model is discretized to obtain the discrete model as shown in Fig. 6. In the figure, S_{cc} is the discrete form of the cooling capacity. S_{js} is the discrete form of fault-related parameters, e.g., the tube inside diameter $d_{tube,in}$ and the fin surface area A_{fin}. Since the physics-based model describe the continuous relationship between cooling capacity and fault-related parameters [18], the "mapping matrices" discretized from the physics-based model represent discrete relationships between S_{cc} and S_{js}.

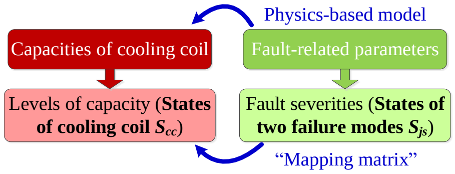
Physics-based model
Capacities of cooling coil
Fault-related parameters
Levels of capacity (**States of cooling coil S_{cc}**)
Fault severities (**States of two failure modes S_{js}**)
"Mapping matrix"

Fig. 6. The relationship between S_{cc} and S_{js}

To discretize the physics-based model, fault-related parameters are sampled randomly. By using Monte-Carlo simulation, N_{smp} sets of ($d_{tube,in}$, A_{fin}) are randomly sampled from fault-related parameters. Two mapping matrices representing relationships between S_{js} and S_{cc} are extracted from these samples. One mapping matrix represents probabilities of S_{js} given S_{cc} and is denoted by $M_{js|cc} = \{m_{ij,js|cc}\}$, with

$$m_{ij,js|cc} = \Pr(S_{js} = j \mid S_{cc} = i) = n_{ij,cc} \big/ n_{i,cc}\,, \tag{22}$$

where $n_{ij,cc}$ is the number of sample points corresponding to $S_{cc} = i$ and $S_{js} = j$, and $n_{i,cc}$ is the number of points belonging to $S_{cc} = i$. Similarly, a mapping matrix $M_{cc|js} = \{m_{ji,cc|js}\}$, with $m_{ji,cc|js} = \Pr(S_{cc} = i \mid S_{js} = j)$, representing probabilities of the cooling coil's states given the joint states, is also calculated.

B. Method of Estimating RULs of a Component or a System under Multiple Failure Modes

In this subsection, the estimate of state at time *t*, the state transition matrix, and parameters of time-duration distributions are estimated by combining those of the two failure modes given mapping matrices $M_{cc|js}$ and $M_{js|cc}$ first. Given these estimates, the method presented in subsection III-C is used to estimate RULs of the cooling coil. Since the two failure modes are independent, it is reasonable to assume that $\Pr(S_{tube}(t)=n)$ and $\Pr(S_{fin}(t) = m)$ are independent, and the probability $\Pr(S_{js}(t) = i)$ is calculated as:

$$\Pr\left(S_{js}(t) = i\right) = \Pr\left(S_{tube}(t) = n\right) \cdot \Pr\left(S_{fin}(t) = m\right), \tag{23}$$

where the joint state $S_{js}(t) = i$ corresponds to $S_{tube}(t) = n$ and $S_{fin}(t) = m$. Given probabilities of each joint state S_{js}, the state S_{cc} of the cooling coil can then be estimated based on the mapping matrix $m_{ij,cc|js}$ as

$$\hat{S}_{cc}(t) = \arg\max_{j} \Pr\left(\sum_{i=0}^{N_{js}-1} m_{ij,cc|js} \cdot \Pr\left(S_{js}(t) = i\right)\right). \tag{24}$$

To estimate the state transition matrix $P_{cc} = \{a_{ij,cc}\}$, where $a_{ij,cc} = P(S_{cc}(t) = j \mid S_{cc}(t-1) = i)$, of the cooling coil, the state transition matrix P_{js} of joint states is required. Since the joint state relies on states of tube fouling and dust on fins, P_{js} is the Kronecker product of P_{tube} and P_{fin} which are state transition matrices of the two failure modes:

$$P_{js} = P_{tube} \otimes P_{fin}. \tag{25}$$

Based on the Bayes rule, the relationship between P_{cc} and P_{js} is obtained as:

$$\begin{aligned}&\sum_{i=0}^{N_{js}-1}\left[\sum_{j=0}^{N_{js}-1} m_{jm,cc|js}(t) \cdot m_{in,cc|js}(t) \cdot \Pr(S_{js}(t-1) = i)\right.\\ &\qquad\left.\cdot \Pr(S_{js}(t) = j \mid S_{js}(t-1) = i)\right]\\ &= \Pr(S_{cc}(t) = m \mid S_{cc}(t-1) = n) \cdot \Pr(S_{cc}(t-1) = n).\end{aligned} \tag{26}$$

Given (26), the elements of the P_{cc} are obtained as:

$$\begin{aligned}&\Pr(S_{cc}(t+1) = m \mid S_{cc}(t) = n)\\ &= \left\{\sum_{i=1}^{N_{js}}\left[\sum_{j=1}^{N_{js}} m_{jm,cc|js}(t+1) \cdot m_{in,cc|js}(t) \cdot \Pr(S_{js}(t) = i)\right.\right.\\ &\qquad\left.\left.\cdot \Pr(S_{js}(t+1) = j \mid S_{js}(t) = i)\right]\right\} \Big/ \Pr(S_{cc}(t) = n),\end{aligned} \tag{27}$$

where $\Pr(S_{js}(t+1)=j|S_{js}(t)=i)$ is the element of P_{js}; $\Pr(S_{js}(t) = i)$ is obtained based on (23); and $\Pr(S_{cc}(t) = n)$ is obtained based on (24). To estimate parameters of time-duration distributions of the cooling coil's HSMM, N_{smp} state sequences $Q_{i,tube}$, $i = 1, \ldots, N_{smp}$ and $Q_{i,fin}$, $i = 1, \ldots, N_{smp}$ of fault severities are generated for tube fouling and dust on fins based on parameters of their HSMMs. Then, N_{smp} joint state sequences $Q_{i,js}$, $i = 1, \ldots, N_{smp}$ are determined based on $Q_{i,tube}$ and $Q_{i,fin}$ since $S_{js} = (S_{tube}, S_{fin})$. After counting time-durations of each joint state in $Q_{i,js}$, the average value of time-durations of State *j* is calculated and denoted by $\bar{D}(S_{js} = j)$. The expected time-duration of State *i* of the cooling coil is then calculated as

$$\bar{D}(S_{cc} = i) = \sum_{j=0}^{N_{js}-1} m_{ji,cc|js} \cdot \bar{D}(S_{js} = j). \tag{28}$$

Thus, the estimated Poisson distribution parameter of State *i* for the cooling coil is obtained as $\hat{\lambda}_i = \bar{D}(S_{cc} = i)$.

To estimate RULs of the air handling system, the method developed above is used. Since both the cooling coil and the supply fan influence the performance of the air handling system, it is reasonable to create the system's HSMM by 1) combining HSMMs of the cooling coil and the decrease in supply fan efficiency; or 2) by combining HSMMs of tube fouling, dust on fins and decrease in supply fan efficiency. The first way needs to discretize physics-based models twice, and the second way only need to discretize the physics-based model once. The second way introduces less information loss, thus is used. Since three failure modes are considered, the system has $4^3 = 64$ joint states. By substituting the physics-based model of fans into the model of cooling coils [18], the cooling capacity is a function of $d_{tube,in}$, A_{fin} and e_{sf}. Two mapping matrices are extracted from this function to estimate RULs of the system in a similar way. Since (24)-(28) combine HSMM parameters of the three failure modes directly rather than calculated iteratively via Gibbs sampling, the method is computationally practical.

V. Experimental Results

Our method is tested against a small simulated building introduced in subsection V-A. To illustrate our methods, three cases are considered, including 1) estimating states of a cooling coil under multiple failure modes in subsection V-B; 3) estimating RUL of the cooling coil in V-C; and 4) estimating RUL of the air handling system in V-D.

A. The Small Simulated Building to Test Our Methods

Therefore, simulation data are used to test our methods. Since the fault-prognosis problem is not sensitive to problem size, a small building with two rooms and a VAV air handling system is considered. By using DesignBuilder [23], the rough building and HVAC structures were established, and were then imported into EnergyPlus [18] to select appropriate component models and parameters to simulate faults. The simple building has two 95.517 m^3 rooms. Most parameters of the HVAC system are set by EnergyPlus automatically based on loads. For instance, under the normal condition, the tube diameter $d_{tube,in}$ is 0.01445 m; the surface area of fins is 43.59555 m^2; and the fan efficiency is 0.7. The system operation is simulated for three years. For the cooling coil, tube fouling is simulated six times by following a Poisson distribution. Each time $d_{tube,in}$ is gradually reduced by 50% and recovers instantly at the end of the time duration. Dust on fins is also simulated six times, and A_{fin} is reduced by 50% each time. Similarly, for the supply fan, a decrease in fan efficiency is simulated five times. To simplify the problem, compound faults are not considered. Simulation data are collected every 10 minutes and are divided into two groups, where 50% of the data is used for training and the rest for testing.

B. Estimate States of a Cooling Coil

For the cooling coil, two failure modes are estimated by using our method as shown in Fig. 7. In the figure, the x-axis is the time and the y-axis is the state of the cooling coil. Actual states and estimated states are represented by black dashed lines and blue stars, respectively. States corresponding to the normal condition, tube fouling, and dust on fins are denoted by '0,' '1' and '2.' In the figure, most actual states and their estimates are the same and are overlapped. However, some false alarms do exist. For instance, there is no fault on 8/17 in the second year, but a dust on fins is falsely detected because of measurement noise. To evaluate the estimated states of HSMMs, *F*-measure is used. This is because *F*-measure is the harmonic mean of precision and recall, namely, it tells how precise the estimator is, as well as how robust it is [24]. Four statistical measures are used to generate *F*-measure, including 1) True Positive (*TP*); 2) True Negative (TN); 3) False Positive (FP); and 4) False Negative (FN). The *F*-measure of Failure Mode *i* or Fault Severity *i* is defined as [24]

$$F\text{-measure}_i = 2TP_\text{i} / (2TP_i + FN_i + FP_i). \tag{29}$$

If false identification rates are 0, then FN_i and FP_i are 0, and F-measure$_i$ is 1 as shown in (29). Thus, *F*-measures is close to 1 if estimates are accurate. *F*-measures of the normal condition, tube fouling, and dust on fins are 0.996, 0.974 and 0.973, indicating that false identification rates are low. The estimation of parameters and states took 1.4 hours and 0.162 seconds, respectively.

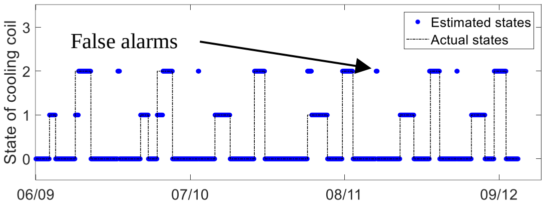

Fig. 7. Estimate failure modes of cooling coil using our method (HSMM)

If the HMM is used, as shown in Fig. 8, there are more false alarms as compared to those using HSMMs. *F*-measures of the normal condition, tube fouling, and dust on fins are 0.993, 0.958 and 0.933, and are worse than those of HSMMs. This is because the geometric distribution is implicitly used in HMMs but this assumption may not be reasonable in practice.

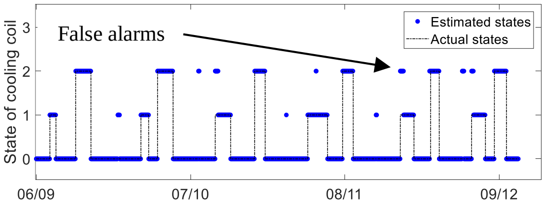

Fig. 8. Estimate failure modes in the cooling coil using an HMM

If the standard Viterbi algorithm is used, it takes 5.6 hours to estimate the parameters of HSMM by using Gibbs sampling, and is much larger than 1.4 hours required by our method. Similarly, it takes 0.736 sec to estimate states of the HSMM by using the Viterbi algorithm, and is larger than 0.162 sec required by our method. Additionally, *F*-measures of the normal condition, tube fouling, and dust on fins are 0.996, 0.969 and 0.975 as shown in Fig. 9, and are approximately the same as those obtained by using our method. This is because all actual state transition points are covered by potential state transition points under consideration.

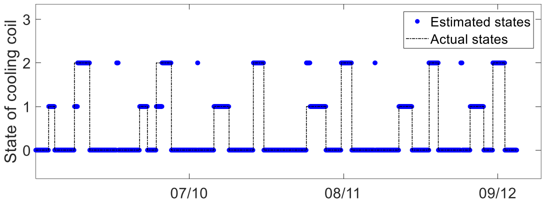

Fig. 9. Estimate failure modes in the cooling coil using the standard Viterbi algorithm (HSMM)

Severities of the cooling coil's failure modes are estimated as shown in Fig. 10. The normal condition is denoted by '0,' and the three fault severities are denoted by '1,' '2,' and '3.' Their *F*-measures are 0.999, 0.965, 0.829 and 0.917. Since *F*-measures are close to one, estimates are accurate.

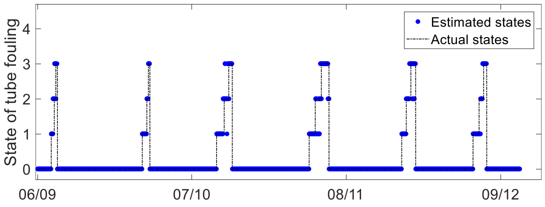

Fig. 10. Estimated severities of tube fouling by using an HSMM

C. Estimate RUL of a Cooling Coil

By using our method presented in subsection III-C, RULs of the cooling coil are estimated in Fig. 11. It can be seen that differences between actual RULs and estimates are small.

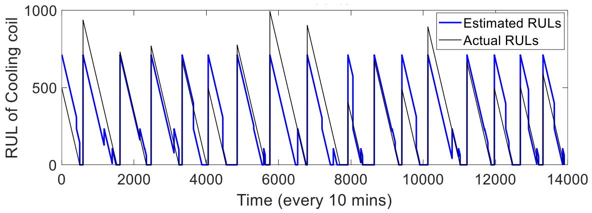

Fig. 11. Actual and estimated RULs of the cooling coil

To evaluate the performance of our fault prognosis method, five performance metrics discussed in [25] are used, including 1) "accuracy" representing the difference between the real RUL and its estimated value; 2) "precision" quantifying the dispersion of prediction errors around its mean; 3) the mean absolute percentage error; 4) the prognostics horizon estimating the time at which the first prediction is within the confidence interval; and 5) the Relative Accuracy (RA) measure permitting one to assess accuracies of estimated RULs at a different time *t*. For the cooling coil, RULs are relatively precise (accuracy equals 0.666). The value of the precision measure is good with a dispersion of around 151.634 time units, or 15.163 hours. The mean absolute percentage error is 65.828. The value of the prognostics horizon is equal to 46 time units, or 4.6 hours. This is small compared to the cooling coil's mean-time-to-failure that is usually several months. RA illustrates the accuracy of estimates at three intervals: the beginning, the middle and the end of data, and is [0.704 0.946 0.745]. Thus, estimates are accurate since they are close to one.

D. Estimate RUL of an Air Handling System

RULs of the air handling system are estimated as shown in Fig. 12 by using our method presented in subsection III-C. In this case, accuracy is 0.615; the precision measure is 197.43 time units, namely 19.743 hour; the mean absolute percentage error is equal to 88.58 time units (8.58 hours); the value of the prognostics horizon is equal to 50 time units (5 hours); and relative accuracies are [0.668 0.835 0.604]. Compared to errors of cooling coils' RUL, errors here are larger. This is because RUL of the air handling system is influenced by more failure modes than those for the cooling coil. The discretization of the physics-based model of the air handling system introduces additional information loss.

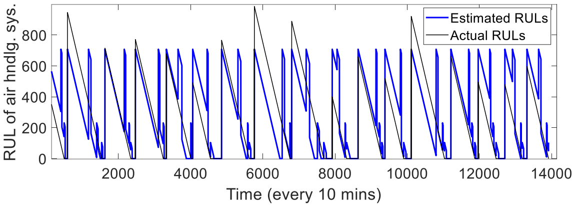

Fig. 12. Actual and estimated RULs of the air handling system

VI. Conclusion

This paper presented an HSMM-based method to estimate the RULs of an air handling system and its components for gradual faults. To estimate parameters and states of HSMMs with low false identification rates and low computational effort, an efficient statistical method is developed. To estimate the RUL of a component or a system, a statistical method is developed to discretize physics-based models to capture the impacts of multiple failure modes. Our method provides a generic framework to estimate the RUL of an air handling system and its components and can be extended to other complex systems.

References

[1] U.S. Department of Energy, *Building Energy Data Book*. Available: http://buildingsdatabook.eren.doe.gov, 2009.

[2] ASHRAE equipment life expectancy chart, Available: chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/http://www.cullum inc.com/wp-content/uploads/2013/02/ASHRAE_Chart_HVAC_Life_Expectancy %201.pdf

[3] Building optimization and fault diagnosis source book, IEA ANNEX 25, 1996.

[4] R. K. Singleton, E. G. Strangas and S. Aviyente, "Extended Kalman filtering for remaining-useful-life estimation of bearings," IEEE Transactions of Industrial Electronics, Vol. 62, No. 3, 2015, pp. 1781-1790.

[5] M. E. Orchard and G. J. Vachtsevanos, "A particle-filtering approach for on-line fault diagnosis and failure prognosis," *Transactions of the Institute of Measurement and Control*, Vol. 31, No. 3-4, 2009, pp. 221-246.

[6] X. H. Jin, Y. Sun, Y. Wang, T. W. S. Chow, "Anomaly Detection and Fault Prognosis for Bearings," *IEEE Transactions on Instrumentation and Measurement*, Vol. 65, No. 9, 2016, pp. 2046-2054.

[7] P. Tse, D. Atherton, "Prediction of machine deterioration using vibration based fault trends and recurrent neural networks," *Transactions of the ASME: Journal of Vibration and Acoustics*, Vol. 121, No. 3, 1999, pp. 355–362.

[8] W. Q. Wang, M. F. Golnaraghi and F. Ismail, "Prognosis of machine health condition using neuro-fuzzy systems," *Mechanical Systems and Signal Processing*, Vol. 18, No. 4, 2004, pp. 813–831.

[9] W. Teng, X. L. Zhang, Y. B. Liu, A. Kusiak and Z. Y. Ma, "Fault Prognosis and Remaining Useful Life Prediction of Wind Turbine Gearboxes Using Current Signal Analysis," *Energies,* Vol. 10, No. 1, 2016, pp. 1-16.

[10] D. A. Tobon-Mejia, K. Medjaher and N. Zerhouni, "CNC machine tool's wear diagnostic and prognostic by using dynamic Bayesian networks," *Mechanical Systems and Signal Processing*, Vol. 28, 2012, pp. 167-182.

[11] H. K. Wang, Y. F. Li, Y. Liu, Y. J. Yang and H. Z. Huang, "Remaining useful life estimation under degradation and shock damage," *Journal of risk and reliability*, Vol. 229, No. 3, 2015, pp. 200-208.

[12] X. Zhang, R. Xu, C. Kwan, S. Y. Liang, Q. Xie and L. Haynes, "An integrated approach to bearing fault diagnostics and prognostics," *in Proceedings of American Control Conference*, Portland, OR, USA, 2005.

[13] M. Dong and D. He, "A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology," *Mechanical Systems and Signal Processing*, Vol. 21, No. 5, 2007, pp. 2248–2266.

[14] D. E. Adams, "Nonlinear damage models for diagnosis and prognosis in structural dynamic systems," in *SPIE Conference Proceedings*, vol. 4733, 2002, pp. 180–191.

[15] J. Luo, K. R. Pattipati, L. Qiao, and S. Chigusa, "Model-based prognostic techniques applied to a suspension system," Transactions on Systems, Man, and Cybernetics, vol. 38, 2003, pp. 1156–1168.

[16] M. Abbas and G. J. Vachtsevanos, "A System-level approach to fault progression analysis in complex engineering systems," *in Proceedings of the Annual Conference of the Prognostics and Health Management Society*, Sans Diego, CA, 2009.

[17] L. R. Rodrigues, "Remaining useful life prediction for multiple-component systems based on a system-level performance indicator," *IEEE Transactions on Mechatronics*, 2017, DOI 10.1109/TMECH.2017.2713722.

[18] EnergyPlus Engineering Reference, http://apps1.eere.energy.gov/buildings/energyplus/pdfs/engineeringreference.pdf.

[19] Y. Yan, P. B. Luh and K. R. Pattipati, "Fault Diagnosis of Components and Sensors in HVAC Air Handling Systems with New Types of Faults," *IEEE Access*, DOI (identifier) 10.1109/ACCESS.2018.2806373.

[20] T. Ryden, "EM versus Markov chain Monte Carlo for estimation of hidden Markov Models: A computational perspective," *Bayesian Analysis.*, Vol. 3, No. 4, 2008, pp. 659–688.

[21] P. M. Djuric and J. H. Chun, "An MCMC sampling approach to estimation of nonstationary hidden Markov model," *IEEE Transactions on Signal Processing*, Vol. 50, No. 5, 2002, pp. 1113 -1123.

[22] J. Sansom and P. J. Thomson, Fitting hidden semi-Markov models, NIWA Technical Report 77, ISSN 1174-2631, 2000.

[23] DesignBuilder 2.1 User Manual, http://www.designbuildersoftware.com/docs/designbuilder/DesignBuilder_2.1_Users-Manual_Ltr.pdf.

[24] T. Mulumba, A. Afshari, K. Yan, W. Shen, and L. K. Norford, "Robust model-based fault diagnosis for air handling units," *Energy Buildings*, Vol. 86, 2015, pp. 698–707.

[25] K. Medjaher, D. Tobon-Mejia and N. Zerhouni, "Remaining useful life estimation of critical components with application to bearings," *IEEE Transactions on Reliability*, Vol. 61, No. 2, 2012, pp. 292-302.