Multi-rate Coupled Hidden Markov Models and Their Application to Machining Tool-Wear Classification

Özgür Çetin,* *Member, IEEE,* Mari Ostendorf, *Fellow, IEEE,* and Gary D. Bernard

***Abstract*— This paper introduces multi-rate coupled hidden Markov models (multi-rate HMMs for short) for multiscale modeling of nonstationary processes, extending traditional HMMs from single to multiple time scales with hierarchical representations of the process state and observations. Scales in the multi-rate HMMs are organized in a coarse-to-fine manner with Markov conditional independence assumptions within and across scales, allowing for a parsimonious representation of both short- and long-term context and temporal dynamics. Efficient inference and parameter estimation algorithms for the multi-rate HMMs are given, which are similar to the analogous algorithms for HMMs. The model is applied to the classification of tool wear in titanium milling, for which acoustic emissions exhibit multiscale dynamics and long-range dependence. Experimental results show that the multi-rate extension outperforms HMMs in terms of both wear prediction accuracy and confidence estimation.**

***Index Terms*— Hidden Markov model, multi-rate hidden Markov model, multiscale statistical modeling, confidence, tool wear, tool-wear monitoring, and milling.**

EDICS Categories: SSP-HIER, MAL-PATT, MAL-APPL

I. Introduction

Machining operations such as milling, turning, and drilling are an indispensable part of manufacturing processes, and losses due to undetected worn tools can be very costly. Machining with a worn tool decreases productivity, and produces reduced quality products. Common industrial practice of replacing cutters at regular intervals can be ineffective due to wide variations in a tool's lifetime; yet this practice does not completely prevent costly failures, and leads to unnecessary delays. Therefore, automatic monitoring systems that can reliably operate across a variety of conditions are highly desired, and have been the focus of significant amount of research [13], [45], [30], [50], [24].

Tool wear monitoring is difficult, because machining processes are highly complex dynamic processes effected by a variety of factors including cutting parameters, instrumentation, and tool and workpiece properties. Analytic methods that can accurately relate the cutting parameters and sensor measurements to the wear amount are not available for the most part, or of limited capability. Therefore, much of the research on automatic wear prediction has focused on data-driven statistical methods. Initial work in this area used static classifiers such as neural networks and Gaussian mixture models (see, e.g., [13], [14], [50] for reviews), which do not model any time-dependent characteristics in the sensory signals. Recent work with HMMs (e.g. [41] and [24] for milling, [27] and [19] for drilling, and [52] for turning) have shown the utility of temporal information. For well behaved materials, HMMs may be sufficiently powerful, but for other materials the wear may be more accurately predicted by exploiting temporal information over multiple time scales, not only the information in a single scale. Wear inherently progresses over multiple time scales, ranging from distinct phases of cutting within a cutting pass (such as entry to the piece, bulk cutting, and exit from the piece) to transients and other short-term phenomena (e.g., due to material removal) occurring at the millisecond level. In particular, acoustic vibrations from titanium milling experience noisy/quiet transient periods that can last up to thirty seconds [40]. Such temporal information can be a good indicator of the wear amount, but existing (static *or* dynamic) monitoring systems do not utilize it. For this purpose, the present paper proposes an extension of HMMs for multiscale modeling, and shows its utility in a difficult titanium milling task. We note that the multi-rate extension developed here is motivated by the milling application, but not specific to it. For example in [6], we have also successfully used multi-rate coupled HMMs for acoustic modeling of speech.

An HMM characterizes a sequence of observations via a hidden state sequence. Roughly speaking, states represent the modes of the nonstationary process, and observations are then emitted from these modes. The strong conditional independence assumptions of HMMs lead to efficient estimation and inference algorithms, but they also limit the kind of phenomena that can be accurately represented by HMMs, in particular multiscale dynamics and long-term dependence. As we elaborate on in subsequent sections, faithfully representing multiple scale-based observation and state sequences in an HMM is difficult if not impossible, and the common workaround of oversampling the coarse-scale observations has increased computational cost as well as limitations in modeling. The use of a tree-based multiscale model, as in [9], captures scale dependencies but the temporal dependencies are only weakly modeled.

Ö. Çetin is with International Computer Science Institute, Berkeley, CA 94704. E-mail: ocetin@icsi.berkeley.edu, Phone: (510) 666–2945, Fax: (510) 666–2956.

M. Ostendorf is with University of Washington, Seattle, WA 98105. E-mail: mo@ee.washington.edu, Phone: (206) 221–5748, Fax: (206) 543–3842.

G. D. Bernard is with Boeing Commercial Airplanes, Seattle, WA 98124. E-mail: gary.d.bernard@boeing.com, Phone: (253) 931–5986, Fax: (253) 931–5987.

Here we extend HMMs to multi-rate coupled HMMs for modeling nonstationary processes with multiscale characteristics, expanding on earlier work reported in [8]. A multi-rate HMM is a collection of scale-based pairs of state and observation sequences, coupled together by sparse stochastic dependencies. Roughly speaking, each pair of state and observation sequences is a scale-based HMM with length inversely proportional to the granularity of the scale. States within each scale form a Markov chain, modeling intra-scale temporal dependencies. Inter-scale dependencies are modeled via tree-structured Markov dependencies between state sequences. Such a judicious choice of dependencies and parameters are particularly advantageous for tool-wear monitoring, where data is scarce due to labeling costs. In addition, the multi-rate model provides better estimates of classifier uncertainty as compared to an HMM, due to the fact the observations from different scales are not modified by oversampling coarser scales to make the rate match the rate of the finest scale. Good confidence prediction is important for tool-wear monitoring, especially if the system is used as an aid to a human operator (vs. as the decision maker).

The organization of this paper is as follows. In Section II, we provide background information about tool-wear monitoring, and briefly review HMMs and previous work in multiscale modeling in Section III. In Section IV, we describe multi-rate coupled HMMs and present their inference and estimation algorithms. Our tool-wear monitoring system architecture is described in Section V, and experimental results comparing HMMs and multi-rate HMMs in a titanium milling task are presented in Section VI. Finally, we conclude in Section VII with a summary and discussion of possible extensions.

II. Tool-wear Monitoring

Tool-wear monitoring refers to the automatic classification of the amount of wear on tools used for metal cutting operations such as milling, turning, and drilling. In this work, we focus on milling. In milling jargon, the metal being machined is referred to as the *workpiece*, and the *tool* is the cutter that physically comes in contact with the workpiece. In a typical milling task, a tool is used for multiple cutting passes, where a *pass* is a single cutting session with defined start and end times.

Tool wear is affected by many factors, referred to here as cutting parameters, including revolutions per minute, feed rate and depth of cut; workpiece properties such as hardness and thermal conductivity; and cutter properties such as tool dimensions and geometry. Depending upon the type of work, varying amounts of wear may be acceptable. Direct measurements of the wear on cutting edges are difficult to obtain, and most tool-wear monitoring systems rely on sensory signals measured from the cutting process to predict wear. Tools used in our milling task are relatively small ($1/2''$ in diameter), and cutting forces, current, and power are not very sensitive to track wear on such small size tools [14], [24]. We use acoustic vibrations instead, as they have been successfully used in similar tasks [41], [1], [23], [40].

Much of the previous work in milling tool-wear prediction has focused on steel with few results reported on titanium [40], [23], in part because titanium poses new challenges. Titanium is chemically reactive, and tends to weld to the cutting edges during cutting, and form a so-called built-up edge in the milling jargon [40]. As the built-up edge volume exceeds a threshold, the bonding forces are overcome by the cutting forces, leading to the chipping of pieces off the primary cutting edges, and premature tool failure. Each welding and release cycle can last from one second to thirty seconds, and the quiet periods are characterized by reduced cutting forces, vibrations, and the rate of wear progression. It has been hypothesized that the primary reason for why the existing systems particularly fail when applied to titanium is that they do not model such long-term characteristics [40]. Among other difficulties associated with titanium are its low thermal conductivity, increasing temperature at the cutter-workpiece interface, and the heterogeneous structure of titanium, containing random hard spots [20].

III. HMMs and Previous Work in multiscale Modeling

HMMs characterize a sequence of observations $\{O_t\}_{t=0}^{T-1}$ via a discrete state sequence $\{S_t\}_{t=0}^{T-1}$:

$$p(\{O_t, S_t\}_{t=0}^{T-1}) \equiv \prod_{t=0}^{T-1} p(S_t|S_{t-1})\, p(O_t|S_t) \tag{1}$$

where S_{-1} is a null start state, $p(S_0)$ is the initial state distribution, $p(S_t|S_{t-1})$ is the state transition matrix, and $p(O_t|S_t)$ is the state-conditional observation distribution. The HMM factorization in Equation 1 involves two independence assumptions: states constitute a first-order Markov chain, and observations are conditionally independent of everything else given their respective states. The states $\{S_t\}$ are hidden, and need to be marginalized out to obtain the probability of observations. The HMM assumptions can be exploited to efficiently perform this and related probabilistic inference tasks with $O(TN^2)$ computational complexity, N being the cardinality of the states, via the forward-backward algorithm. Similarly, the maximum likelihood parameter estimation of HMMs in the presence of the hidden variables can be efficiently done via the Baum-Welch algorithm, an instance of the generic expectation-maximization (EM) algorithm [15]. (See [46], [18] for tutorials that further describe HMMs and the associated inference and estimation algorithms.) The popularity of HMMs partially stems from the availability of these algorithms, but the independence assumptions availing these algorithms also restrict the phenomena, in particular multiscale dynamics and long-term dependence, that can be efficiently represented by the HMMs.

There are a few limitations of the HMM framework for multiscale modeling. First, HMMs represent the process state with a single state variable, and representing multi-component state spaces requires assigning a unique state to each composite state configuration. The resulting state space is exponential in the number of components, increasing both the computational cost of inference and number of free parameters. Second, HMMs assume a single observation sequence at one time scale. To map a multiscale process to a single scale, one

needs to either oversample the coarser observation sequences, which introduces redundancy and gives artificially biased class *a posteriori* probabilities during classification, or downsample the finer ones, which discards useful information. Third, HMMs propagate context through an information bottleneck, the current state, which can encode at most $\log_2 N$ bits of information due to the Markov chain structure. As the time lag between the past and future increases, the information between the past and future that HMM represents decays exponentially fast [3]. Structural constraints in the state transitions, such as a left-to-right transition topology, can be used to propagate context information to some degree, but this approach is ineffective for long sequences.

To address these and other deficiencies of HMMs, various extensions of HMMs with multiple state and observation sequences have been proposed in the literature. We group these extensions based on whether the sequences are time-synchronous (so that the modeling is essentially single scale) or multiscale. Coupled HMMs [5], mixed memory models [48], and pairwise Markov chains [43] involving multiple pairs of state and observation sequences that depend on each other through hidden states, are well-known examples from the first group. A related model is the factorial HMM [25], in which there is a single observation sequence, but multiple state sequences that indirectly interact via their common influence on observations. These models have found wide use in automatic speech recognition for multi-stream [4], [39] and audio-visual modeling [36]. multiscale statistical models in the second group have been explored in many different facets of signal processing and data fusion; see [53] for an extensive review. Some representative examples include the hidden Markov trees [12], multi-resolution autoregressive models [2], multiscale Kalman filters [10], multiscale Markov random fields [35], and multiresolution hidden Markov chains [21]. Two multiscale models that are closely related to the multi-rate coupled HMMs proposed here are the two-dimensional multi-resolution HMMs [34], and the hierarchical HMMs [22], [38]. Both models involve scale-based state and observation sequences, organized in a hierarchical manner from coarser to finer scales. After we introduce the multi-rate coupled HMMs, we will give a detailed comparison between these models and the multi-rate coupled HMMs in Section IV-C.

IV. Multi-rate Coupled HMMs

Multi-rate coupled HMMs are a multiscale generalization of HMMs for the *joint* modeling of scale-based observation sequences. Similar to HMMs, hidden states are assumed to underlie the observations, but unlike HMMs, there are multiple state sequences, each of which is associated with a particular scale. Naturally, such scale-based sequences are of different lengths. We use this multiscale hidden state structure to parsimoniously represent both the intra-scale and inter-scale dependencies.

In a K-rate coupled HMM, there are K scales, each of which is associated with an observation sequence $\{O^k_{t_k}\}^{T_k-1}_{t_k=0}$, and a state sequence $\{S^k_{t_k}\}^{T_k-1}_{t_k=0}$. Here k indexes the scales, and T_k is the length of observation sequence at the scale k. The scales are ordered from the coarsest $k=1$ to the finest $k=K$. We assume that the scale k is M_k times faster than the scale $k-1$, i.e. $T_k = M_k\, T_{k-1}$, and hence the term multi-rate coupled HMMs. The K-rate coupled HMMs characterize these K scale-based state and observation sequences through the factorization

$$p(\{O^1_{t_1}, S^1_{t_1}\}^{T_1-1}_{t_1=0}, \ldots, \{O^K_{t_K}, S^K_{t_K}\}^{T_K-1}_{t_K=0}) \equiv \prod_{k=1}^{K} \prod_{t_k=0}^{T_k-1} p(S^k_{t_k}|S^k_{t_k-1}, S^{k-1}_{\lfloor t_k/M}$$

where S^k_{-1} is a null start state for the k-th scale, and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x (hence $\lfloor t_k/M_k \rfloor$ is the index of the unique observation in the scale $k-1$ overlapping with the observation t_k in the next-coarsest scale k). We assume that the states are discrete, and the observations are continuous. Thus, the parameters of a K-rate HMM consist of the initial state probabilities $p(S^k_0|S^{k-1}_0)$, the state-transition probabilities $p(S^k_{t_k}|S^k_{t_k-1}, S^{k-1}_{\lfloor t_k/M_k \rfloor})$, and the parameters associated with the state-conditional observation distributions $p(O^k_{t_k}|S^k_{t_k})$. The latter parameters are not tied across scales, but the transition probabilities are assumed to be time-homogeneous.

The sparse factorization in Equation 2 involves three main conditional independence assumptions: first, conditioned on the states at the previous-coarsest scale $k-1$ ($\{S^{k-1}_{t_{k-1}}\}$), the states at the scale k ($\{S^k_{t_k}\}$) constitute a first-order Markov chain; second, conditioned on the overlapped state at the previous-coarsest scale $k-1$ ($S^{k-1}_{\lfloor t_k/M_k \rfloor}$) and the previous state at the same scale k ($S^k_{t_k-1}$), the state $S^k_{t_k}$ at the scale k is independent of all coarser states starting from the scale $k-1$; and third, observations are independent of everything else given their states. These independence assumptions can be illustrated via the use of graphs. In Figure 1a, a dynamic Bayesian network representation of a 2-rate coupled HMM is given. (Dynamic Bayesian networks are a kind of probabilistic graphical models, using acyclic directed graphs. See e.g., [42], [33], [11] for a mathematically exact interpretation of these graphs as probabilistic models.) Each node in this directed graph corresponds to a random variable on the left-hand side of Equation 2, and the directed parent-to-child arrows correspond to the conditional distributions on the right-hand side of Equation 2. In addition to being a useful visualization tool, Bayesian networks provide generic algorithms for inference, estimation, and decision making, which we will utilize in the coming sections.

As easily seen from Figure 1a or Equation 2, a K-rate coupled HMM essentially amounts to K scale-based HMMs which have been coupled via their states, and roughly put, there is a first-order Markov structure across scales. Dropping $S^{k-1}_{\lfloor t_k/M_k \rfloor}$ from the right hand side of $p(S^k_{t_k}|S^k_{t_k-1}, S^{k-1}_{\lfloor t_k/M_k \rfloor})$ in Equation 2 would render the scale-based sequences independent of each other. While the first-order Markov dependencies across time model the scale-dependent evolution within each scale, the dependencies across scales model coupling between the scale-based parts. The result is a parsimonious model with judicious parameterization, which can represent multiscale dynamics and long-term dependency, and learn such properties from small amounts of data. As the coupling between the

scale-based parts gets stronger, it becomes more beneficial to represent inter-scale dependencies, and the multi-rate HMM-based approaches become more advantageous over the HMM-based approaches.

A. Probabilistic Inference

The states in multi-rate HMMs are hidden, and the probabilistic inference in multi-rate HMMs involves the calculation of various marginal and *a posteriori* probabilities in the presence of these hidden variables. The marginal probability of observations $p(\mathcal{O})$ where $\mathcal{O} \equiv \cup_k \{O_{t_k}^k\}_{t_k=0}^{T_k-1}$, is needed for evaluating models against data, and the state posteriors such as $p(S_{t_k}^k|\mathcal{O})$ are needed in the EM-based parameter estimation routines. The calculation of these probabilities by brute force summation is prohibitively expensive, but can be efficiently done by exploiting the sparse factorization in Equation 2, via the generic graphical modeling algorithms: the junction tree algorithm and the message passing algorithm [42].

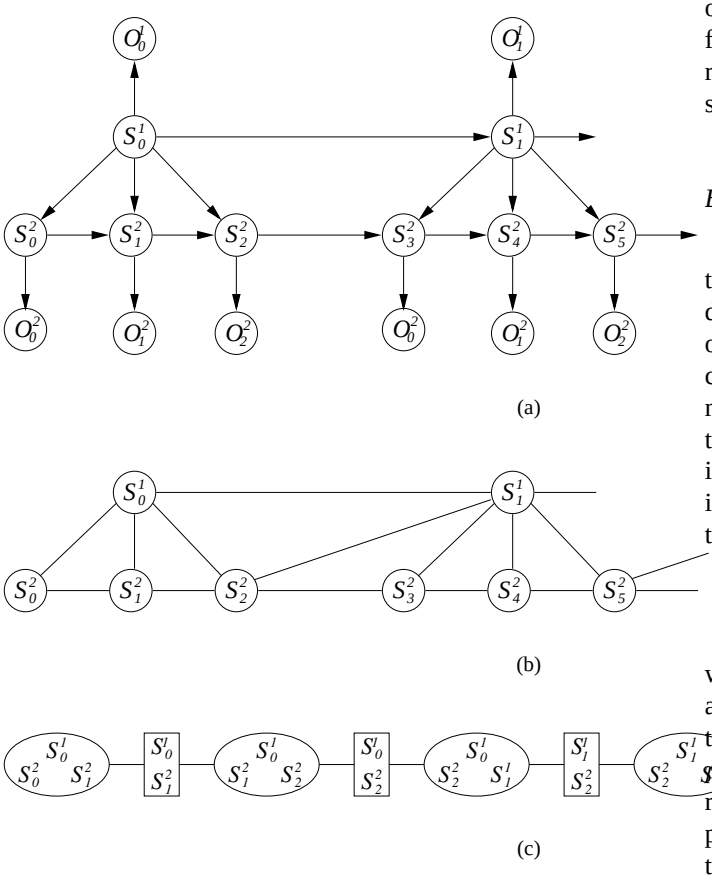

Fig. 1. (a) The Bayesian network illustration of a 2-rate HMM with $M_2 = 3$ (M in the figure), and $T_1 = T$; (b) the corresponding moralized graph, which happens to be already triangulated (observations are removed to simplify display); and (c) the corresponding junction tree for the triangulated graph. In (c), the ovals represent cliques, and the rectangles represent clique separators which are the intersections of neighboring cliques. The messages are first sent forwards from the root (the leftmost node) towards the leave (the rightmost node), and then backwards.

First, the junction tree algorithm manipulates the original directed graph into a hypergraph by a series of graph transformation operations: moralization (where an edge is added between every pair of unconnected parents, and the directionality of arrows is dropped), triangulation (where extra edges are added to the moralized graph so that there are no chordless cycles with length greater than four), and organization of the cliques of the triangulated graph into a clique tree with the running intersection property.[1] These steps are illustrated in Figure 1 for a 2-rate HMM. Second, the message passing algorithm uses this clique tree as a data structure with potential functions defined over its cliques. (Potentials are derived from the parent-to-child conditional factors of the original probability factorization.) Similar to the HMM forward-backward algorithm, the messages are exchanged between the neighboring cliques to ensure the global consistency of the clique potentials. The various marginal and *a posteriori* probabilities are then extracted from the modified potential functions. The most-likely hidden state configuration can also be obtained using a particular message passing protocol [11]. For a K-rate HMM, the computational cost of the resulting algorithm is of $O(T_K N^{K+1})$, assuming that the cardinalities of the sample spaces of all state chains are equal to N for illustration purposes. The inference equations for the 2-rate coupled HMMs that we use in our tool-wear monitoring system are given in Appendix I.

B. Parameter Estimation

The parameters of the multi-rate HMM consist of initial state and state transition probabilities, and observation distributions, which we assume are modeled by mixtures of Gaussian components. We use the maximum likelihood criterion to estimate these parameters from data. The direct maximization of the likelihood function is intractable due to hidden states, and we resort to the EM algorithm for its iterative maximization. The EM algorithm replaces the incomplete likelihood by the *expected* complete likelihood. At the i-th iteration the parameters are updated according to

$$\theta^{(i)} = \operatorname*{argmax}_{\theta} \mathbf{E}\left[\log p_\theta(\mathcal{S}, \mathcal{O}) \,|\, \mathcal{O}, \theta^{(i-1)}\right] \tag{3}$$

where $\theta^{(i)}$ denotes the parameter estimate at the i-th iteration, and $\mathcal{S}$ is a completion of unobserved states. The expectation is taken with respect to the hidden state *a posteriori* probabilities $p_{\theta^{(i-1)}}(\mathcal{S}|\mathcal{O})$, which can be obtained by the inference algorithm described in Section IV-A using the parameters from the previous iteration. The sequence of parameters obtained from this procedure is guaranteed to monotonically not decrease the incomplete likelihood function, and converge to a local maximum of it [15]. Maximizing Equation 3 is straightforward, and results in updates similar to the Baum-Welch updates for estimating HMMs [46]. The update equations for the 2-rate coupled HMMs are given in Appendix II.

[1] A clique $\mathcal{C}$ is a maximal complete subgraph. Any two vertices in $\mathcal{C}$ are connected, and $\mathcal{C}$ is not strictly contained in any other such subgraph. A *chord* is an edge between two nonconsecutive vertices in a cycle. A set of cliques has the running intersection property when, if a vertex appears in two cliques then it appears in every clique along the path joining these cliques [33].

C. Comparison to HMMs and to Previous Work in multiscale Modeling

In theory, an HMM can represent any state-space model by an appropriate grouping of state and observation sequences of that model. The conversion of a multi-rate HMM to an HMM can be done in two alternative ways. In the first approach, one oversamples all coarser-scale state sequences to make them synchronous with the finest-scale state sequence, and then takes their Cartesian product. Next, the observation distributions of the HMM are specially arranged so that not every product state emits observations for all scales at every time step; the product state emits an observation at a particular scale commensurate with the corresponding scale of the multi-rate HMM. The resulting model is equivalent to the original multi-rate HMM, but not really an HMM due to the special emitting structure. The second approach is a widely used workaround in the speech recognition literature when multiscale observations are employed—the so-called feature concatenation approach. Feature concatenation simulates multi-rate HMM only approximately. In this approach, in addition to state oversampling, observation sequences are oversampled to make all sequences synchronous with each other, their Cartesian product within each time slice is taken, and the resulting meta observation sequence is modeled by a factored observation distribution. Oversampling in feature concatenation introduces redundancy, which in turn results in overconfident posterior probabilities during classification, because roughly speaking, the evidence from the coarser scales is counted multiple times. In either approach, the product state space is exponential in the number of scales. The product state space results in an exponentially higher cost of inference. For a K-rate HMM, the inference in the emulating HMM costs $O(TN^{2K})$, as compared to $O(TN^{K+1})$ when the inference is performed using the multi-rate HMM inference algorithm. In addition, the state transitions and observation distributions need to be specially parameterized to simulate the state transition and observation distributions of the multi-rate HMM. If not, the number of parameters is exponentially more, and learning generalizing relationships from sparse data becomes more difficult.

The multi-rate HMMs have similarities to and differences from the multiscale models proposed in the literature. Here, we will mainly focus on the 2D multi-resolution HMMs [34] and the hierarchical HMMs [22], which too are instances of dynamic Bayesian networks (see, e.g., [38]), and most similar to multi-rate HMMs. Like the 2D multi-resolution HMMs [34] and the hierarchical HMMs [22], the multi-rate HMMs characterize inter-scale dependencies via tree-structured coarse-to-fine dependencies among the scale-based state sequences. Similar to the two-dimensional multi-resolution HMMs but unlike hierarchical HMMs, the alignment between the scale-based sequences in multi-rate HMMs is formulated deterministically, but the extension to probabilistic alignment is straightforward. Unlike either model, each scale-based state variable sequence in multi-rate coupled HMMs constitutes a Markov chain when conditioned on coarser ancestor variables. The coupling of state chains in the multi-rate HMM is also similar to the coupled HMMs [4], [5], [16], [48], [39], [43], which use inter-stream Markov dependencies to couple two same-rate hidden Markov models. However, coupled HMMs are single rate, so they also introduce redundancy from oversampling coarser scales.

V. System Architecture

Our tool-wear monitoring system models the progress of wear from sharp to worn over multiple cutting passes during the lifetime of a tool. Both the gross characteristics of wear from one pass to another, as well as the characteristics within a pass are modeled. The overall system consists of three modules: feature extraction, statistical modeling, and decision making. In feature extraction, cepstral features are extracted from the acoustic vibrations. Let $\{\mathcal{O}_l, W_l\}_{l=1}^{L}$ be the feature and wear, respectively, sequences associated with a tool that has lasted L passes. We assume that there is a single wear level associated with each cutting pass, which is not overly restrictive given that multiple categories $\mathcal{W}$ are being used. In statistical modeling, the progress of wear over multiple cutting passes is modeled as a meta HMM:

$$p(\{\mathcal{O}_l, W_l\}_{l=1}^{L}) \equiv p(W_1)\, p(\mathcal{O}_1|W_1) \prod_{l=2}^{L} p(W_l|W_{l-1})\, p(\mathcal{O}_l|W_l). \quad (4)$$

The wear process within a pass is modeled by the term $p(\mathcal{O}_l|W_l)$, which we characterize using either an HMM or a multi-rate HMM. In decision making, the meta HMM is used to estimate the wear posterior probabilities for each cutting pass in an online manner, which can either be used directly to make predictions, or further refined by a second-stage classifier. Details associated with each module follow.

A. Feature Extraction

Short-term transients due to chipping and material removal occur at the millisecond level. To capture these and other short-term characteristics, a sequence of cepstral features are computed at the flute frequency, extracting 20 cepstral coefficients based on a 25-th order linear prediction analysis [54]. The analysis window size is set to 110% of the time between flute strikes, since partially overlapping windows reduce boundary effects [23]. Only the 1-st, 4-th, 5-th, and 6-th cepstral coefficients are kept to reduce dimensionality, since they were found to have significantly higher discriminatory information based on the scatter ratios of coefficients coming from sharp and worn passes, $\rho \equiv (\mu_0 - \mu_1)^2/(\sigma_0^2 + \sigma_1^2)$, where (μ_0, σ_0) and (μ_1, σ_1) denote the sample mean and variance calculated from cutting passes labeled sharp and worn, respectively.[2] Alternatively, linear discriminant analysis or principal component analysis could be used for reducing dimensionality (see Section VI-D).

Acoustic vibrations, in particular those from titanium milling, exhibit long-term behavior characterized by the frequency and type of short-term transients. To capture these, an

[2]The features from partially worn cutting passes are not used in this analysis in order to simplify the feature selection problem in a binary classification setting.

average of short-time cepstral features over every 10 rotations is taken (hence over every 40 frames, because tools in our database have four flutes). A window size of 40 frames was suggested by the expert knowledge, but this choice has been verified experimentally against the other window sizes. Windows are nonoverlapping, so the resulting long-term feature sequence is 40 times slower than the original short-term one.

To minimize the effects of tool-specific biases and sensor effects, cepstral mean subtraction has been applied on a tool-by-tool basis to both short- and long-term cepstral features. Cepstral mean subtraction is a standard technique in speech recognition to reduce inter-speaker variability and convolutive noise [54], and has also been found useful for tool-wear monitoring [23]. A cepstral mean vector for each tool is calculated by taking the sample mean of cepstral features coming from the first three passes of that tool, if they are available. If not, the global mean of such first three passes across all tools is used as the cepstral mean. This cepstral mean is subtracted from all features coming from that tool.

B. Statistical Modeling

The progress of wear is hierarchically modeled at two levels. A wear-level HMM characterizes the wear dynamics from one cutting pass to another (cf. Equation 4), with states that correspond to the wear levels, W_l, ordered according to increasing amounts of wear. Within this HMM, the observations are modeled by another HMM or multi-rate HMM. The *a posteriori* probability $p(W_l|\mathcal{O}_1,\ldots,\mathcal{O}_l)$ is obtained by normalizing the forward probabilities $\alpha_l(w) \equiv p(W_l = w, \mathcal{O}_1,\ldots,\mathcal{O}_l)$ in the recursion

$$\alpha_l(w) = \sum_{w'\in\mathcal{W}} \alpha_{l-1}(w')\, p(W_l = w|W_{l-1} = w')\, p(\mathcal{O}_l|W_l = w). \quad (5)$$

The state transitions $p(W_l|W_{l-1})$ are constrained to have a left-to-right transition topology to reflect the fact that wear cannot decrease over time. The observation probability $p(\mathcal{O}_l|W_l = w)$ characterizes observations from the l-th pass under wear w, using the corresponding HMM or multi-rate HMM.

When using an HMM to model $p(\mathcal{O}_l|W_l = w)$, we can use either short- or long-term cepstral features, or their concatenation as observations. Alternatively, multi-rate HMMs can be used for modeling $p(\mathcal{O}_l|W_l = w)$. We use a 2-rate HMM with the short- and long-term cepstral features in the fine and coarse scales, respectively, aimed at capturing noisy/quiet transient regimes due to the built-up edge formation in titanium milling. For both the HMM and multi-rate HMM process models, fully-connected state transition topologies are chosen based on pilot experiments, the absence of any prior knowledge suggesting a left-to-right topology (unlike steel milling [24]), and the observation that noisy/quiet regimes can occur more than once in a cutting pass. Also for both the HMM and multi-rate HMM process models, the observation distributions are modeled by mixture of diagonal Gaussians, and the number of states and mixture components are determined via cross-validation (cf. Section VI).

The recursion in Equation 5 is dominated by the observation probabilities involving thousands of observations, and the state transition probabilities contribute little to the sum. To reduce this mismatch, it was useful to normalize the observation probabilities by the length of that sequence in the log domain, similar to the weighting of a language model in the speech recognition systems [54].

C. Decision Making

Having observed the observation sequences for the cutting passes up to pass l, the Bayes optimal decision for the wear category of the l-th pass is given by

$$\lambda_l(\mathcal{O}_1,\ldots,\mathcal{O}_l) \equiv \operatorname*{argmax}_{w\in\mathcal{W}} p(W_l = w|\mathcal{O}_1,\ldots,\mathcal{O}_l). \quad (6)$$

Equation 6 is invoked for each pass of a tool, using the full sequence observations up to the current pass rather than starting from the previous decision. As such, the detected wear sequence is not required to be monotonic, and the algorithm can recover from bad decisions given additional evidence.

Previous work [49], [24] has found that a generalized linear model (GLM) [37] can further improve the original posterior estimates from Equation 5, by modifying them as follows:

$$\hat{p}(W_l = w|\mathcal{O}_1,\ldots,\mathcal{O}_l) \equiv \frac{\exp(\beta_w^\top \eta_l)}{\sum_{w'\in\mathcal{W}} \exp(\beta_{w'}^\top \eta_l)} \quad (7)$$

where $\top$ denotes transposition, β_w is the regression coefficients, and $\eta_l \equiv [1\ \log\alpha_l(1)\ \cdots\ \log\alpha_l(|\mathcal{W}|)]^\top$ is the GLM features. Using a GLM for posterior correction is important because the posterior estimates of Equation 5 tend to be skewed towards zero or one. Such skewed posteriors render operating at different points of the receiver operating characteristics curve ineffective for trading missed detections for false alarms, as illustrated in [23]. As we will show in Section VI-D, GLMs help recover from overconfident posteriors with only a few additional parameters. GLMs also provide a convenient framework for classifier combination by using η_l's from different classifiers as features.

VI. Experiments

A. Data

The data are from aerospace milling machines cutting titanium blocks, and collected by Boeing Commercial Airplane Group using a numerically controlled machining center with the cutting parameters given in Table I. Fresh tools were used to climb-cut notches for multiple cutting passes until they become worn. An accelerometer was mounted radially on the front plate of the spindle housing, recording wide-band machining vibrations due to chip-formation and transients caused by edge breakdown. At the end of the selected passes, a human operator measured the amount of wear on the cutting edges.

The wear measurements were quantized into three nonoverlapping categories, A, B, and C, corresponding roughly to the tools that are sharp (less than $0.0024''$ of wear), worn but still usable (between $0.0024''$ and $0.0049''$ of wear), and worn (more than $0.0049''$ of wear), respectively. Not all cutting passes have their wears measured, because measurements required dismantling the machining center. For some

TABLE I
CUTTING PARAMETERS

Parameter	
Block size	18”
Rockwell-C hardness	32-34
RPM	733
Feed rate (in/min)	14.7
Axial depth-of-cut	1.0”
Radial depth-of-cut	0.2”
Tool diameter	0.5”
Wear endpoint	0.005”

TABLE II
TRAINING AND TESTING DATA STATISTICS

Data Set	*# of tools*	*# of passes*	*# of labeled passes*		
			A	*B*	*C*
Training	6	70	18	12	9
Testing	9	75	15	17	10

unlabeled passes, the assumption that wear cannot decrease over time can be used to infer the wear category from the measurements forward and backward in time. We collectively refer to the labels from the hand measurements and the labels deduced as such, as the true labels. For other unlabeled passes, neighboring passes rule out only one category, but such partial information is incorporated into training via the EM algorithm. Previous work on steel milling showed that such full EM training with unlabeled data is more efficient than the Viterbi-style approaches [24].

We divide the data into training and test sets on a tool-by-tool basis so that the training and testing tools are nonoverlapping. The various statistics for the training and testing sets are given in Table II.

B. System Training

The parameter estimation is done according to the maximum likelihood criterion. The initial wear-level probabilities and the wear-level transition probabilities of the meta HMM, $p(W_l)$ and $p(W_l|W_{l-1})$ in Equation 4, are estimated by fitting a Poisson model to the labeled training passes, and calculating the initial state and state-transition probabilities from this model. The wear-dependent cutting pass models, $p(\mathcal{O}_l|W_l = w)$ for $w \in \mathcal{W}$ in Equation 4, whether they be modeled by HMMs or multi-rate HMMs, are jointly estimated using the EM algorithm. Here the EM algorithm operates at two levels, automatically handling both the missing wear labels, and the hidden states in the HMM or multi-rate HMM pass models. The GLM regression coefficients, β_w in Equation 7, are estimated from the labeled cutting passes using the iterative reweighted least squares algorithm [37] by leave-two-out cross-validation. In leave-k-out cross-validation, k tools are set aside for testing while the remaining tools are used for estimating the system parameters, which are in turn used to obtain the features, η_l in Equation 7, for the k tools that were set aside. By rotation, the features for all training tools are obtained, and the iterative reweighted least squares algorithm is invoked with these features. This cross-validation scheme is necessary to avoid optimizing the GLM coefficients on the same data as the cutting pass models are optimized on.

C. Figures of Merit

We use the error rate and normalized cross-entropy (NCE) for classifier comparison. Each cutting pass contributes as one example. Error rates are accompanied with p-values to assess the significance of the performance differences. The p-values reported here are exact according to a Binomial model, since our sample size may not be large enough for the asymptotic Gaussian approximation to hold [26].

NCE evaluates the accuracy of the class *a posteriori* probability estimates [51]. The NCE of a classifier which predicts class $C \in \mathcal{C}$ from features X is given by

$$NCE \equiv \frac{H(C) - H(C|X)}{H(C)} \tag{8}$$

where $H(C)$ and $H(C|X)$ are the empirical *a priori* and *a posteriori*, respectively, entropies:

$$H(C) \equiv -\frac{1}{T}\sum_{t=1}^{T} \log p(C_t), \quad H(C|X) \equiv -\frac{1}{T}\sum_{t=1}^{T} \log p(C_t|X_t)$$

where (C_t, X_t) denote the t-th testing pair in a test set of size T. In the NCE metric, an incorrect classification decision that is made with a high *a posteriori* probability incurs more cost than the one made with a lower *a posteriori* probability. Ignoring features and using solely the prior results in an NCE of zero; one is the perfect score; and highly confident incorrect decisions results in negative NCE. For numerical stability, we constrain the probability estimates to be between ϵ and $1 - \epsilon$ with $\epsilon = 10^{-10}$. NCE results are accompanied with the significance levels at which they are different from zero, according to a Wilcoxon rank-sum test comparison of $\{\log p(C_t|X_t)\}$ and $\{\log p(C_t)\}$ [28].

The NCE effectively evaluates the confidence of the classifier decision. Confidence prediction is important for tool-wear monitoring, because usually a human uses the information from the monitoring system to make the final tool replacement decisions, and the system is not the sole decision maker.

D. Results

Five sets of models are compared: the HMMs using short- and long–term features (HMM–FINE and HMM–COARSE, respectively), the HMM using the concatenated short- and long-term features (HMM–BOTH), the classifier combination of HMM systems using short- and long-term features (HMM–BOTH+), and the 2-rate HMM using the short- and long-term features (MHMM). The HMM system using the concatenated short- and long-term features is similar to the other HMM systems except that it uses the stacked short- and long-term features after oversampling the long-term ones. The classifier combination of the HMM systems is done via a GLM, using the log-probability features coming from these systems. Feature concatenation and classifier combination of HMMs are two very different alternatives to the multi-rate HMM for integrating short- and long-term information. In the feature concatenation approach, the integration is done at the state

TABLE III

THE PARAMETERS FOR THE HMM AND 2-RATE HMM SYSTEMS, DETERMINED VIA CROSS-VALIDATION ON THE TRAINING SET. THE PAIRS IN THE MHMM ROW DENOTE THE RESPECTIVE PARAMETERS FOR THE SHORT- AND LONG-TERM SCALE CHAINS.

Model	*# of states*	*# of mixtures*	*# of parameters*
HMM–FINE	2	2	37
HMM–COARSE	2	2	37
HMM–BOTH	2	2	69
MHMM	2/4	2/1	95

TABLE IV

THE ERROR RATES AND p-VALUES FOR STATISTICAL DIFFERENCE FROM THE A PRIORI CLASSIFIER, FOR THE HMM AND 2-RATE HMM SYSTEM ON THE TRAINING SET VIA CROSS-VALIDATION (CV), AND ON THE TESTING SET WITH AND WITHOUT THE GLM POSTERIOR CORRECTION.

Model	*Training Set w/ CV*		*Testing Set*		*Testing Set w/ GLM*	
	error %	*p-value*	*error %*	*p-value*	*error %*	*p-value*
HMM–FINE	64	N/S	62	N/S	43	$2 \cdot 10^{-8}$
HMM–COARSE	8	10^{-9}	43	$2 \cdot 10^{-3}$	43	$2 \cdot 10^{-3}$
HMM–BOTH	8	10^{-9}	45	$5 \cdot 10^{-3}$	43	$2 \cdot 10^{-3}$
HMM–BOTH+	N/A	N/A	N/A	N/A	45	$5 \cdot 10^{-3}$
MHMM	62	N/S	62	N/S	33	$2 \cdot 10^{-8}$

TABLE V

THE NCES AND p-VALUES FOR THE STATISTICAL DIFFERENCE FROM AN NCE OF ZERO, FOR THE HMM AND 2-RATE HMM SYSTEMS OF TABLE IV. ITALIC p-VALUES INDICATE NCE VALUES THAT ARE SIGNIFICANTLY WORSE THAN ZERO.

Model	*Training Set w/ CV*		*Testing Set*		*Testing Set w/ GLM*	
	NCE	*p-value*	*NCE*	*p-value*	*NCE*	*p-value*
HMM–FINE	-0.05	N/S	$3 \cdot 10^{-3}$	N/S	-0.54	N/S
HMM–COARSE	0.61	10^{-10}	-1.34	N/S	-1.82	N/S
HMM–BOTH	0.64	10^{-10}	-1.39	N/S	-1.65	N/S
HMM–BOTH+	N/A	N/A	N/A	N/A	-3.42	$9 \cdot 10^{-3}$
MHMM	-0.01	N/S	$5 \cdot 10^{-3}$	N/S	0.23	$7 \cdot 10^{-3}$

level on the order of milliseconds, whereas in the classifier combination approach, the integration is done at the cutting-pass level on the order of minutes. The concatenation approach models the scale dependence by oversampling and potentially overweighting the coarse scale, whereas classifier combination ignores inter-scale dependence but avoids overweighting. In contrast, the multi-rate HMM is able to represent the scale dependence without overweighting the coarse scale.

The cross-validated number of states and number of mixture components per observation distribution for each system is given in Table III. We note that even though the 2-rate system has more parameters than the HMM systems have, the HMM systems with equal or more parameters have performed worse than the ones in Table III during cross-validation. No results with GLMs on the training set are reported, because the GLM regression coefficients are estimated on the cross-validation data (cf. Section VI-B). The classification error rates of the HMM systems and the 2-rate HMM system are in Table IV, with p-values denoting the significance level at which the corresponding classifier is different from the *a priori* classifier; p-values that are more than 0.10 are reported as not significant (N/S). For reference, the error rate of the *a priori* classifier which always outputs the most-frequent wear category in the training set, is 54% on the training set, and 66% on the testing set. The NCEs are in Table V, with p-values denoting the significance level at which the NCE of the corresponding system is different from zero.

We make following observations based on Tables IV and V. First, without the GLM, the performances of the HMM system using short-term features and the 2-rate HMM system on the training and testing sets are almost identical in terms of both the error rate and the NCE; their error rates are almost identical to that of the *a priori* classifier; their NCE scores are near zero; and therefore they are not statistically different from the *a priori* classifier. However, with the GLM, while both the HMM system using short-term features and the 2-rate HMM system significantly reduce error rate over the *a priori* classifier, the reduction is largest for the 2-rate HMM system (yet their error rates are statistically different only at the level $p = 0.14$ due to the small sample size). However in terms of NCE, the 2-rate HMM system outperforms the HMM system at the level $p = 0.06$. Among all systems, the 2-rate HMM system with the GLM is the only system with a positive NCE that is significantly different from zero. Second, the performances of the HMM system using long-term features and the HMM system using concatenated short- and long-term features, without the GLM, are identical in terms of both the error rate and the NCE; they obtain excellent performance on the training set; and they are statistically different from the *a priori* classifier in terms of both criteria. However, their NCEs are negative on the testing set, indicating poor confidence prediction. For these two HMM systems, the GLM improves neither error rate nor NCE. Third, we observe that the classifier combination of the HMM systems using short- and long-term features performs similar to the other HMM systems in terms of error rate, but is significantly worse than all other systems in terms of NCE, with a statistically inferior NCE. The poor performance of the GLM combination of the HMM systems using short- and long-term features also shows that a scaling of scores from different sources is not sufficient to deal with the problem of redundancy in oversampled features.

In other experiments not reported here for brevity, we obtained similar results to those in Tables IV and V, when different number of features (2-8), or alternative dimensionality reduction schemes (linear discriminant analysis or principle component analysis) were used. The HMM-based feature concatenation and classifier combination approaches were consistently unsuccessful at integrating short- and long-term information, and the resulting classifiers were uninformative with negative NCE values, while the 2-rate HMM approach giving the most successful results with positive NCE.

Overall, neither the feature concatenation nor the classifier combination approach is satisfactory for combining short- and long-term information, especially in terms of NCE. This finding confirms the importance of efficient modeling of titanium wear processes at multiple time scales without introducing redundancy. Either of the HMM-based combination approaches results in large negative NCE values (indicative of the poor posterior probability estimation), and the wear

posterior probabilities are not reflective of the true amount of wear. The 2-rate HMM approach seems to be the most promising approach and achieves the best classification and confidence performance when used in combination with a GLM. Thus, indeed there are gains from joint characterization of short- and long-term features in a multiscale modeling framework, and the GLM posterior correction is effective.

VII. Summary and Extensions

In summary, the multiscale extension of HMMs efficiently characterizes stochastic processes with multiscale dynamics, using sparse temporal and coarse-to-fine dependencies to represent and learn the joint multiscale dynamics and long-term dependence. Multi-rate HMMs differ from HMMs in the representation of state-observation sequence pairs at more than one time scale. The parsimonious structure of the multi-rate HMMs allows for learning relatively complex models from sparse data. The experiments on a difficult titanium tool-wear monitoring task showed that the multi-rate HMMs perform better than the HMMs and the alternative HMM-based combination approaches utilizing multiscale feature sequences. In particular, we found that the multi-rate HMM with GLM posterior correction gives a better indication of the confidence of its decisions, as measured by positive NCE. Although error rates of the 2-rate HMM system were around 30%, the rates are lower than those of HMM-based systems and the good results in NCE indicate that multi-rate HMMs can provide useful outputs to human operators.

There are a number of natural extensions to the multi-rate HMM framework introduced here. First, the basic multi-rate HMM framework is rather restrictive because of the limited form of dependencies assumed between scales. There may exist residual dependencies among observation sequences, or between the state and observation sequences from different scales, not captured by indirect state coupling. There are more sophisticated cross-scale observation and state dependency modeling techniques (e.g. [6], [43], [44]) that could also be effective for multi-rate HMMs. Second, in our analysis, we assumed that every coarse-scale observation corresponds to a constant number of finer-scale ones, i.e. the rate of the model is time-invariant. As opposed to such a fixed-rate analysis, a variable-rate signal analysis can tailor the analysis to focus more on information-bearing regions with transitions or transients [6]. Third, the parameter estimation of the multi-rate coupled HMMs in this paper is done according to the maximum likelihood criterion. For classification problems, the discriminative criteria such as maximum conditional likelihood [17], or minimum classification error rate [31] can be more appropriate. Multi-rate HMMs and tool-wear monitor systems can also be parameterized conditionally and trained as such, as in conditional random fields [32], to focus on the prediction of wear categories from observations. As we have demonstrated by using GLMs for secondary posterior correction, there could also be additional benefits from combining multi-rate HMMs with discriminative approaches, such as using the Fisher score of a multi-rate HMM in a kernel machine [29]. Fourth, in the tool-wear application, the signal processing used for multiscale feature extraction is fairly simple, and more advanced multiscale signal analysis techniques such as wavelets, could be more beneficial. In addition, the multi-rate HMMs are particularly well suited for the multi-sensor approach, especially if sensors require different rate analysis.

We note that an important limitation of our work and previous tool-wear research has been the very little amount of data available for system development and performance assessment. Sparse data prohibit sophisticated models that might better characterize complex wear processes, but also require more data for estimation. Sparse data also avoid comparing models to high certainty and extracting diagnostics information. We believe that the tool-wear monitoring research can benefit a lot from large, carefully collected databases, as it has been the case in speech recognition, with which we drew many parallels.

VIII. Acknowledgments

This work was supported by the Office of Naval Research (ONR-2883401). The views and conclusions in this paper are those of the authors, and do not necessarily reflect the views of the sponsors. This work was performed while the first author is at University of Washington. Initial experiments with the multi-rate HMM appeared in [23] for a different milling task, which involved a collaboration of the authors with R. K. Fish. The authors thank Dr. Fish for useful discussions, and Boeing Commercial Airplanes Manufacturing R&D group for sharing their machining expertise and providing titanium data.

References

[1] L. Atlas, M. Ostendorf, and G.D. Bernard. Hidden Markov models for monitoring machine tool wear. In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 3887–3890, 2001.

[2] M. Basseville, A. Benveniste, K.C. Chou, S.A. Golden, R. Nikoukhah, and A.S. Willsky. Modeling and estimation of multiresolution stochastic processes. *IEEE Trans. on Information Theory*, 38(2):766–784, March 1992.

[3] Y. Bengio and P. Frasconi. Diffusion of context and credit information in Markovian models. *Journal of Artificial Intelligence Research*, 3:249–270, 1995.

[4] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. of Int. Conf. on Spoken Language Processing*, pages 426–429, 1996.

[5] M. Brand and N. Oliver. Coupled hidden Markov models for complex action recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[6] Ö. Çetin. *Multi-rate Modeling, Model Inference, and Estimation for Statistical Classifiers*. Ph.D. thesis, University of Washington, 2004.

[7] Ö. Çetin and M. Ostendorf. Multi-rate hidden Markov models for monitoring machining tool wear. Technical Report UWEETR-2004-0011, University of Washington Department of Electrical Engineering, 2003.

[8] Ö. Çetin and M. Ostendorf. *Multi-rate hidden Markov models and their application to machining tool-wear classification*. In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 837–840, 2004.

[9] K.C. Chou and L.P. Heck. A multiscale stochastic modeling approach to the monitoring of mechanical systems. In *Proc. of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 25–27, 1994.

[10] K.C. Chou, A.S. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization. *IEEE Trans. on Automatic Control*, 39:464–478, 1994.

[11] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

[12] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Signal Processing*, 46(4):886–902, 1998.

[13] L. Dan and J. Mathew. Tool wear and failure monitoring techniques for turning–A review. *Intl. J. of Machine Tools and Manufacture*, 30(4):579–598, 1990.

[14] Snr. D.E. Dimla. Sensor signal for tool-wear monitoring in metal cutting operations–A review of methods. *Intl. J. of Machine Tools and Manufacture*, 40:1073–1098, 2000.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B*, 39:185–197, 1977.

[16] S. Dupont and H. Bourlard. Using multiple time scales in a multi-stream speech recognition system. In *Proc. of Eurospeech*, pages 3–6, 1997.

[17] D. Edwards and S.L. Lauritzen. The TM algorithm for maximizing a conditional likelihood function. *Biometrika*, 88:961–972, 2001.

[18] Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Trans. on Information Theory*, 48:1518–1569, 2002.

[19] H.M. Ertunc, K.A. Loparo, and H. Ocak. Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs). *Intl. J. of Machine Tools and Manufacture*, 41:1363–1384, 2001.

[20] E.O. Ezugwu and Z.M. Wang. Titanium alloys and their machinability–A review. *Intl. J. of Materials Processing Technology*, 68:262–274, 1997.

[21] L. Fouque, A. Appriou, and W. Pieczynski. Multiresolution hidden Markov chain model and unsupervised image segmentation. In *Proc. of 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 121–125, 2000.

[22] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[23] R.K. Fish. *Dynamic Models of Machining Vibrations, Designed for Classification of Tool Wear*. PhD thesis, University of Washington, 2001.

[24] R.K. Fish, M. Ostendorf, G.D. Bernard, and D.A. Castanon. Multilevel classification of milling tool wear with confidence estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(1):75–85, 2003.

[25] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.

[26] L. Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 532–535, 1989.

[27] L.P. Heck and J.H. McClellan. Mechanical system monitoring using hidden Markov models. In *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1697–1700, 1991.

[28] R.V. Hogg, A.T. Craig and J. McKean. *Introduction to Mathematical Statistics*. Prentice Hall, 2004.

[29] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1999.

[30] E. Jantunen. A summary of methods applied to tool condition monitoring in drilling. *Intl. J. of Machine Tools and Manufacture*, 42:997–1010, 2002.

[31] B.H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing*, 40:3043–3053, 1992.

[32] J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of Proc. Intl. Conf. on Machine Learning*, 591–598, 2001.

[33] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[34] J. Li, R.M. Gray, and R.A. Olshen. Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models. *IEEE Trans. on Information Theory*, 46(5):1826–1841, 2000.

[35] M.R. Luettgen, W.C. Karl, and A.S. Willsky. Multiscale representations of Markov random fields. *IEEE Trans. on Signal Processing*, 41:3377–3396, 1993.

[36] J. Luettin, G. Potamianos, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 169–172, 2001.

[37] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. New York Chapman and Hall, 1989.

[38] K.P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[39] H.J. Nock and M. Ostendorf. Parameter reduction schemes for loosely coupled HMMs. *Computer Speech and Language*, 17:233–262, 2003.

[40] M. Ostendorf, L. Atlas, R. Fish, Ö. Çetin, S. Sukittanon, and G.D. Bernard. Joint use of dynamical classifiers and ambiguity plane features. In *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, pages 3589–3592, 2001.

[41] L. Owsley, L. Atlas, and G. Bernard. Self-organizing feature maps and hidden Markov models for machine-tool monitoring. *IEEE Trans. on Signal Processing*, 45:2787–2798, 1997.

[42] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1991.

[43] W. Pieczynski. Pairwise Markov chains. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:634–639, 2003.

[44] W. Pieczynski and F. Desbouvries. On triplet Markov chains. In *Proc. of Intl. Symp. on Applied Stochastic Models and Data Analysis*, 2005.

[45] P.W. Prickett and C. Johns. An overview of approaches to end milling tool monitoring. *Intl. J. of Machine Tools and Manufacture*, 39:105–122, 1999.

[46] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

[47] R. Shachter. Bayes-Ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proc. of Uncertainty in Artificial Intelligence*, pages 480–487, 1998.

[48] L.K. Saul and M.I. Jordan. Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simple ones. *Machine Learning*, pages 1–11, 1998.

[49] H. Shivakumar. Prediction of tool wear likelihood for milling applications. Master's thesis, Boston University, 2000.

[50] B. Sick. On-line and indirect tool wear monitoring in turning with artificial neural networks: A review of more than a decade of research. *Mechanical Systems and Signal Processing*, 16(4):487–546, 2002.

[51] M. Siu and H. Gish. Evaluation of word confidence for speech recognition systems. *Computer Speech and Language*, 13:299–319, 1999.

[52] L. Wang, M.G. Mehrabi, and Jr. E. Kannatey-Asibu. Hidden Markov model-based tool wear monitoring in turning. *J. of Manufacturing Science and Engineering*, 124(3):651–658, 2002.

[53] A.S. Willsky. Multiresolution Markov models for signal and image processing. *Proc. of the IEEE*, 90(8):1396–1458, 2002.

[54] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.2)*. Cambridge University, 2002.

APPENDIX I
INFERENCE IN THE 2-RATE COUPLED HMMS

In this section we present an inference algorithm for 2-rate coupled HMMs. The inference equations for higher order multi-rate HMMs can be derived analogously [6]. We set $T_1 = T$ and $T_2 = TM$ (M is the coarse-to-fine scale ratio), and define $\mathcal{O} \equiv \{O_t^1\}_{t=0}^{T-1} \cup \{O_\tau^2\}_{\tau=0}^{TM-1}$, and $\mathcal{S} \equiv \{S_t^1\}_{t=0}^{T-1} \cup \{S_\tau^2\}_{\tau=0}^{TM-1}$. We will use the shorthands $X_{<t} \equiv \{X_\tau\}_{\tau=0}^{t-1}$ and $X_{\leq t} \equiv \{X_\tau\}_{\tau=0}^{t}$, and use $A \perp\!\!\!\perp B \,|\, C$ to denote that A is conditionally independent of B given C, i.e. $p(A, B|C) = p(A|C)\, p(B|C)$. Similar to the HMM inference algorithm, the inference in the 2-rate HMM consists of a forward and a backward pass.

A. Forward Pass

We define the forward variables $\alpha_t^1(i,j) \equiv p(S_t^1 = i, S_{tM-1}^2 = j, \{O_\tau^1\}_{\tau=0}^t, \{O_\tau^2\}_{\tau=0}^{tM-1})$, and $\alpha_t^2(i,j) \equiv p(S_{\lfloor t/M \rfloor}^1 = i, S_t^2 = j, \{O_\tau^1\}_{\tau=0}^{\lfloor t/M \rfloor}, \{O_\tau^2\}_{\tau=0}^t)$. The forward recursion proceeds as follows:

for $t = 0$ to $T-1$ **do**
  $\alpha_{tM}^2(i,j) = p(O_{tM}^2 | S_{tM}^2 = j) \sum_k \alpha_t^1(i,k)\, p(S_{tM}^2 = j | S_{tM-1}^2 = k, S_t^1 = i)$
  for $\tau = tM+1$ to $(t+1)M-1$ **do**
    $\alpha_\tau^2(i,j) = p(O_\tau^2 | S_\tau^2 = j) \sum_k \alpha_{\tau-1}^2(i,k)\, p(S_\tau^2 = j | S_{\tau-1}^2 = k, S_t^1 = i)$
  end for

6: $\alpha^1_{t+1}(i,j) = p(O^1_{t+1}|S^1_{t+1} = i) \sum_k \alpha^2_{(t+1)M-1}(k,j)\, p(S^1_{t+1}=i|S^1_t=k)$
7: **end for**

Step 2 is replaced by $\alpha^2_0(i,j) = p(S^2_0=j|S^1_0=i)\, p(O^1_0|S^1_0=i)\, p(O^2_0|S^2_0=j)$ for $t=0$, by using the definition of $\alpha^1_t(i,j)$; Step 6 is not invoked for $t=T-1$. Steps 2 and 4 follow from the 2-rate HMM conditional independence properties (CI1) $O^2_t \perp\!\!\!\perp \mathcal{S}\backslash\{S^2_t, O^2_t\} \mid S^2_t$, and (CI2) $S^2_t \perp\!\!\!\perp \{S^1_{<\lfloor t/M\rfloor-1}, S^2_{<t-1}, O^1_{\leq\lfloor t/M\rfloor}, O^2_{\leq t-1}\} \mid \{S^1_{\lfloor t/M\rfloor}, S^2_{t-1}\}$, whereas Step 6 follows from (CI3) $O^1_t \perp\!\!\!\perp \mathcal{S}\backslash\{S^1_t, O^1_t\} \mid S^1_t$, and (CI4) $S^1_t \perp\!\!\!\perp \{S^1_{<t-1}, S^2_{<tM}, O^1_{<t}, O^2_{<tM}\} \mid S^1_{t-1}$. These properties can be directly verified using Equation 2, or 'read off' from Figure 1a using the Bayes-ball algorithm [47].

B. Backward Pass

We define the backward variables $\beta^1_t(i,j) \equiv p(\{O^1_\tau\}_{\tau=t+1}^{T-1}, \{O^2_\tau\}_{\tau=tM}^{TM-1} | S^1_t = i, S^2_{tM-1} = j)$ and $\beta^2_t(i,j) \equiv p(\{O^1_\tau\}_{\tau=\lfloor t/M\rfloor+1}^{T-1}, \{O^2_\tau\}_{\tau=t+1}^{TM-1} | S^1_{\lfloor t/M\rfloor} = i, S^2_t = j)$. The backward recursion proceeds as follows:

1: **for** $t = T-1$ to 0 **do**
2: $\beta^2_{(t+1)M-1}(i,j) = \sum_k p(S^1_{t+1}=k|S^1_t=i)\, p(O^1_{t+1}|S^1_{t+1}=k)\, \beta^1_{t+1}(k,j)$
3: **for** $\tau = (t+1)M-2$ to tM **do**
4: $\beta^2_\tau(i,j) = \sum_k \beta^2_{\tau+1}(i,k)\, p(S^2_{\tau+1}=k|S^2_\tau=j, S^1_t=i) p(O^2_{\tau+1}|S^2_{\tau+1}=k)$
5: **end for**
6: $\beta^1_t(i,j) = p(S^2_{tM}=k|S^2_{tM-1}=j, S^1_t=i)\, p(O^2_{tM}|S^2_{tM}=k) \beta^2_{tM}(i,k)$
7: **end for**

Step 2 is replaced by the initialization $\beta^2_{TM-1}(i,j) \equiv 1$ for $t = T-1$; Step 6 is not invoked for $t=0$. Step 2 follows from (CI2), (CI4), and (CI5) $\{O^1_{>t+1}, O^2_{>(t+1)M-1}\} \perp\!\!\!\perp \{S^1_t, O^1_{t+1}\} \mid \{S^1_{t+1}, S^2_{(t+1)M-1}\}$, whereas Steps 4 and 6 follow from (CI1), and (CI6) $\{O^1_{>\lfloor t/M\rfloor}, O^2_{>t}\} \perp\!\!\!\perp \{S^2_{t-1}, O^2_t\} \mid \{S^1_{\lfloor t/M\rfloor}, S^2_t\}$. Again, (CI5) and (CI6) can be easily verified by using the multi-rate HMM factorization in Equation 2.

C. Marginal Likelihood and Hidden State Posteriors

The marginal likelihood of observations $\mathcal{O}$ can be calculated from any of the forward variables, for example, $p(\mathcal{O}) = \sum_{i,j} \alpha^2_{TM-1}(i,j)$, which directly follows from the definition of α^2_t. The various state *a posteriori* probabilities can be obtained from the forward and backward variables using the aforementioned independence properties (CI1–CI6) as follows:

$$
\begin{aligned}
p(S^1_{t-1}=i, S^1_t=j, S^2_{tM-1}=k|\mathcal{O}) &\propto \alpha^2_{tM-1}(i,k)\, p(S^1_t=j|S^1_{t-1}=i)\, p(O^1_t|S^1_t=j)\, \beta^1_t(j,k) \\
p(S^1_t=i, S^2_{tM-1}=j, S^2_{tM}=k|\mathcal{O}) &\propto \alpha^1_{t-1}(i,j)\, p(S^2_{tM}=k|S^2_{tM-1}=j, S^1_t=i)\, p(O^2_{tM}|S^2_{tM}=k) \\
&\quad \times \beta^2_{tM}(i,k) \\
p(S^1_t=i, S^2_{tM+k-1}=j, S^2_{tM+k}=k|\mathcal{O}) &\propto \alpha^2_{tM+k-1}(i,j)\, p(S^2_{tM+k}=k|S^2_{tM+k-1}=j, S^1_t=i) \\
&\quad \times p(O^2_{tM+k}|S^2_{tM+k}=k)\, \beta^2_{tM+k}(i,k)
\end{aligned}
\tag{9}
$$

APPENDIX II
PARAMETER ESTIMATION OF THE 2-RATE COUPLED HMMS

In this section we present an EM algorithm for the parameter estimation of 2-rate coupled HMMs. To simplify the presentation, we assume that the observation distributions are represented by single Gaussians, but the extension to the mixture case is straightforward [7], [6]. The parameters in this setting consist of the initial state probabilities, $\pi^1_i \equiv p(S^1_0=i)$ and $\pi^2_{ij} \equiv p(S^2_0=i|S^1_0=j)$, state-transition probabilities, $a^1_{ij} \equiv p(S^1_t=i|S^1_{t-1}=j)$ and $a^2_{ijk} = p(S^2_t=i|S^2_{t-1}=j, S^1_{\lfloor t/M\rfloor}=k)$, and the Gaussian means and covariances of the observation distributions, $p(O^1_t=o^1|S^1_t=i) \equiv \mathcal{N}(o^1; \mu^1_i, \Sigma^1_i)$, and $p(O^2_t=o^2|S^2_t=i) \equiv \mathcal{N}(o^2; \mu^2_i, \Sigma^2_i)$.

Let us denote the EM auxiliary function in Equation 3 by $Q(\theta;\tilde{\theta}) \equiv \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(\mathcal{S},\mathcal{O})$ where we removed the conditioning inside the expectation to simplify the notation. We manipulate $Q(\theta;\tilde{\theta})$ by plugging in $p_\theta(\mathcal{S},\mathcal{O})$ from Equation 2:

$$
\begin{aligned}
Q(\theta;\tilde{\theta}) &= \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(S^1_0) + \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(S^2_0|S^1_0) \\
&\quad + \sum_{t=1}^{T-1} \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(S^1_t|S^1_{t-1}) + \sum_{t=1}^{TM-1} \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(S^2_t|S^1_{\lfloor} \\
&\quad + \sum_{t=0}^{T-1} \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(O^1_t|S^1_t) + \sum_{t=0}^{TM-1} \mathbf{E}_{\mathcal{S}|\mathcal{O},\tilde{\theta}} \log p_\theta(O^2_t|S^2_t) \\
&= \sum_i \gamma^1_0(i) \log \pi^1_i + \sum_{i,j} \gamma^{12}_0(j,i) \log \pi^1_{ij} \\
&\quad + \sum_{t=1}^{T-1} \sum_{i,j} \gamma^{11}_t(j,i) \log a^1_{ij} + \sum_{t=1}^{TM-1} \sum_{i,j,k} \gamma^{122}_t(k,j,i) \log a^1_{ijk} \\
&\quad - \frac{1}{2} \sum_{t=0}^{T-1} \sum_i \gamma^1_t(i) \left(d_1 \log(2\pi) + \log|\Sigma^1_i| + (O^1_t - \mu^1_i)^\top (\Sigma^1_i)^{-1} \right. \\
&\quad - \frac{1}{2} \sum_{t=0}^{TM-1} \sum_i \gamma^2_t(i) \left(d_2 \log(2\pi) + \log|\Sigma^2_i| + (O^2_t - \mu^2_i)^\top (\Sigma^2_i)^{-1} \right.
\end{aligned}
$$

where d_1 and d_2 are the dimensionality of O^1_t and O^2_t, respectively, and the *a posteriori* probabilities $\gamma^1_t(i) \equiv p(S^1_t=i|\mathcal{O})$, $\gamma^2_t(i) \equiv p(S^2_t=i|\mathcal{O})$, $\gamma^{12}_t(j,i) \equiv p(S^1_{\lfloor t/M\rfloor}=j, S^2_t=i|\mathcal{O})$, $\gamma^{11}_t(j,i) \equiv p(S^1_{t-1}=j, S^1_t=i|\mathcal{O})$, and $\gamma^{122}_t(k,j,i) \equiv p(S^1_{\lfloor t/M\rfloor}=k, S^2_{t-1}=j, S^2_t=i|\mathcal{O})$ are easily obtained from Equations 9–11. (It is implicit in our notation that these probabilities are calculated using the parameters $\tilde{\theta}$.) It is now straightforward to maximize Q with respect to θ by differentiation. The updates for the initial state and state-transition probabilities are given by

$$
\pi^1_i \propto \gamma^1_0(i), \quad \pi^1_{ij} \propto \gamma^{12}_0(j,i), \quad a^1_{ij} \propto \sum_{t=1}^{T-1} \gamma^{11}_t(j,i), \quad a^1_{ijk} \propto \sum_{t=1}^{TM-1} \gamma^{122}_t
$$

The updates for the coarse-scale Gaussian means and covariances are given by

$$
\mu^1_i = \frac{\sum_{t=0}^{T-1} \gamma^1_t(i)\, O^1_t}{\sum_{t=0}^{T-1} \gamma^1_t(i)}, \quad \Sigma^1_i = \frac{\sum_{t=0}^{T-1} \gamma^1_t(i)\, O^1_t\, (O^1_t)^\top}{\sum_{t=0}^{T-1} \gamma^1_t(i)} - \mu^1_i\, (\mu^1_i)^\top. \tag{10}
$$

The updates for the fine-scale Gaussian parameters take a similar form, with obvious modifications. (11)