# MULTI-RATE HIDDEN MARKOV MODELS AND THEIR APPLICATION TO MACHINING TOOL-WEAR CLASSIFICATION

*Özgür Çetin, Mari Ostendorf*

Signal, Speech, and Language Interpretation Laboratory
Department of Electrical Engineering, University of Washington, Seattle
{cozgur,mo}@ee.washington.edu

**ABSTRACT**

This paper introduces a multi-rate hidden Markov model (multi-rate HMM) for multi-scale stochastic modeling of non-stationary processes. The multi-rate HMM decomposes the process variability into scale-based components, and characterizes both the intra-scale temporal evolution of the process and the inter-scale interactions. Applying these models to the machine tool-wear classification problem in a titanium milling task shows that multi-rate HMMs outperform HMMs in terms of both accuracy and confidence of predictions.

## 1. INTRODUCTION

Automatic detection of tool wear in metal cutting operations is an important industrial problem, and losses due to undetected worn tools can be very costly. Common industrial practice of replacing at regular intervals can be ineffective and inefficient due to the wide variations in a cutter's lifetime. Therefore, automatic wear prediction systems that can reliably operate across a variety of cutters and cutting conditions are of practical importance.

Direct measurement of physical wear amount on cutting edges is difficult to obtain during cutting, and tool-wear monitoring systems rely on sensory signals measured from the cutting process to predict wear amount using a statistical classifier. Wear progression is a dynamic process, and the importance of temporal information for wear prediction has been demonstrated in previous work, where hidden Markov models (HMMs) were used for modeling both gross characteristics of wear over multiple cutting passes [10, 8] and structure of short-term transients at the millisecond level [16]. However, an HMM models stochastic processes evolving at only one time scale, with limited ability to represent long-term context. Hence, HMMs are not well suited for modeling the wear process for tasks involving materials such as titanium, which exhibit long-term dependence and multi-scale dynamics. In titanium milling, cutters experience cycles of welding and release of titanium to cutting edges, and the time-frequency structure of associated machining vibrations and transients differ considerably from those in the absence of such cycles.

In this work, we introduce multi-rate HMMs for multi-scale stochastic modeling of non-stationary processes, and apply them to the tool-wear monitoring problem. Multi-rate HMMs are a generalization of HMMs in that multiple state and observation sequences are used to represent temporal and observational variability, respectively, at multiple time scales. Scales are organized in a hierarchical manner from coarse to fine, allowing for efficient representation of both short- and long-term context simultaneously. While the tree-structured dependencies across scales characterize the coarse-to-fine inter-scale dynamics, statistical dependencies across time characterize the intra-scale temporal evolution within each scale-based part.

The benefit of modeling of the wear process at multiple time scales through multi-rate HMMs is three fold. First, machining vibrations have both short- and long-term characteristics, and the multi-rate HMM can provide a more accurate description of the wear process by multi-scale modeling. Second, tool-wear system design is constrained by the high costs of data collection, so the amount of training data available for parameter estimation is limited. In this case, using the parsimonious representation of the multi-rate HMM is a key advantage. Third, the classifier confidence is an important consideration in tool-wear monitoring systems, where they are used as an indicator by a human operator, not as a decision maker. The multi-rate HMM classifiers potentially have better confidences, because the wear process model does not make over-simplistic assumptions or introduce any redundant information via over-sampling of multi-scale observation sequences.

In the following sections, we provide a discussion of HMMs and their use in tool-wear monitoring as well as their limitations for stochastic modeling of multi-scale processes. The multi-rate HMM extension, with inference and estimation algorithms, is described next, with experimental results comparing HMMs and multi-rate HMMs in a titanium milling task.

## 2. HIDDEN MARKOV MODEL

An HMM characterizes the joint distribution of a length $T$ time series, $\{o_t\}$, called observations, through an underlying hidden state sequence, $\{s_t\}$,

$$p(\{o_t\},\{s_t\}) \equiv \prod_{t=0}^{T-1} p(s_t|s_{t-1})p(o_t|s_t)$$

in which it is assumed that the state sequence is first-order Markov, $s_{-1}$ is a null start state, and observations are conditionally independent of everything else given their respective states [15]. Roughly speaking, non-stationary behavior of the time evolution is captured by the discrete state sequence, while the consecutive observations produced by the same state captures piecewise stationarity. The HMM's independence assumptions can be exploited to construct a computationally efficient $O(TN^2)$ probabilistic inference algorithm, $N$ denoting the state space cardinality.

The HMM has been used to model non-stationary phenomena ranging from weather patterns to speech acoustics. HMMs have

also been proposed for milling [16, 14, 8], drilling [10, 5], and turning [18]. The deficiency of HMMs for capturing both short- and long-term dynamics for titanium milling has been noted in [14] where ambiguity plane features (a type of time-frequency analysis) were used for capturing local feature dynamics.

Even though the HMM's conditional independence assumptions produce a simple yet powerful model with an efficient inference algorithm, those assumptions also restrict the phenomena that can be parsimoniously represented by the HMMs. First, the assumption that observations are conditionally independent given the hidden state sequence rarely holds, and the state-based dependence is weak. Information between the past and present observations as represented by an HMM decays to zero exponentially fast as the time lag between them increases, due to the underlying Markov chain structure [1]. Though structural constraints in the HMM's state topology can be used to propagate long-term context information to some degree, this approach is not very effective for long observation sequences. Second, an HMM represents the process state with a single state variable and enforces synchrony in state transitions associated with different components of the observation vector. Many processes consist of multiple interacting parts, each of which might evolve at a different time scale. Representation of composite state structures by an HMM requires assigning a unique state to each possible composite state configuration, resulting in an exponential state space which increases both the computational cost of inference and the number of free parameters. Moreover, representation of multi-scale observation sequences by a single observation sequence requires over-sampling of coarser-scale observation sequences to make them synchronous with finer scale observation sequences, resulting in artificially skewed class posterior estimates and over-confident classification decisions due to counting the same evidence multiple times.

## 3. MULTI-RATE HIDDEN MARKOV MODEL

### 3.1. Model Assumptions

The multi-rate HMM is a generalization of the HMM to multiple time scales. The process state is factored into scale-based components, with multiple observation sequences each of which is at a particular resolution of time. In a $K$-rate HMM, the process is modeled at $K$ time scales, and associated with each scale is a hidden state sequence, $\{s^k_{t_k}\}$, and an observation sequence, $\{o^k_{t_k}\}$, $k$ denoting the scale level. Scales are organized in a hierarchical manner from the coarsest $k = 1$ to the finest $k = K$, and the $k$-th scale is $M_k$ times faster than the $(k-1)$-th scale, i.e. $T_k = M_k T_{k-1}$ for $k > 1$ where $T_k$ denotes the length of observation sequence associated with the $k$-th scale. A multi-rate HMM characterizes both the time evolution within each scale-based part and the interactions among them. The joint distribution of state and observation sequences is modeled as,

$$p(\{o^1_{t_1}\}, \{s^1_{t_1}\}, \ldots, \{o^K_{t_K}\}, \{s^K_{t_K}\}) \equiv \prod_{k=1}^{K} \prod_{t_k=0}^{T_k-1} p(s^k_{t_k} | s^k_{t_k-1}, s^{k-1}_{\lfloor t_k/M_k \rfloor}) p(o^k_{t_k} | s^k_{t_k}) \quad (1)$$

where $s^k_{-1}$ is a null start state for the $k$-th scale, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x$ and hence $\lfloor t_k/M_k \rfloor$ is the index of the time frame in the $(k-1)$-th scale covering the $t_k$-th observation in the $k$-th scale. A graphical model illustration of the multi-rate HMM is in Figure 1.

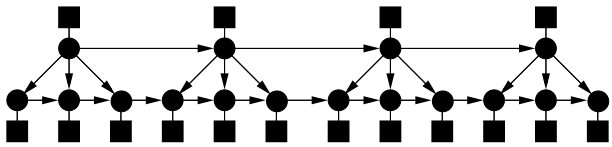

**Fig. 1**. Graphical model illustration of a multi-rate HMM with $K = 2$ and $M_2 = 3$, with the coarse scale at the top. States and observations are depicted as circles and squares, respectively.

In the multi-rate HMM, statistical dependencies across time characterize the temporal dynamics of the scale-based components, whereas the hierarchical dependencies across scale characterize the interaction between the components. Dependencies across time are first-order Markov; those across scale are tree structured. The $K$-rate HMM essentially involves $K$ multi-length HMMs, which are coupled via dependency of state transitions at one scale on the overlapping state variable at the next higher scale. Notice that dropping $s^{k-1}_{\lfloor t_k/M_k \rfloor}$ from the right hand side of the first term in Equation 1 would render the pairs of scale-based observation and state sequences independent of each other.

The use of a factored state representation in which different components can change their states at different time scales ameliorates the representation of context problem. Coarser scale state and observation sequences provide long-term context information to finer scale ones characterizing short-term behavior. State and observation factoring is also used in variations of HMMs for single-rate processes, including factorial HMMs [9], mixed memory models [17] and coupled HMMs [2, 13], in part to reduce the number of free parameters (for more robust estimation) and in part to allow asynchrony between state processes associated with different observation streams. The multi-rate state factoring benefits from these advantages, but also is better suited for modeling long-term context and reducing feature redundancy.

Two-dimensional multi-resolution HMMs [11] and hierarchical HMMs [6] are examples of multi-scale models involving scale-based state and observation sequences. Like the multi-rate HMM, these models use tree-structured coarse-to-fine dependencies to characterize inter-scale dependencies, but they are more restrictive in terms of the assumptions they make: state and observation variables at a given scale are conditionally independent of their distant relatives given their parent or ancestor state variables, and the state sequence at a given scale is disconnected and not a Markov chain.

### 3.2. Probabilistic Inference

The state sequences, $\{s^k_{t_k}\}$, are hidden, so they must be marginalized out when evaluating models against data. The parameter estimation of multi-rate HMMs require the a posteriori probabilities of states given observations. The calculation of these and other quantities by brute force summation of Equation 1 is prohibitively expensive, but they can be efficiently calculated via a generalization of the HMM forward-backward algorithm to the multi-rate HMM. The algorithm exploits Markov independence relationships of the multi-rate HMM to calculate hidden state posteriors given all observation sequences. Forward and backward steps recursively calculate the conditional probability of the current states given past observations, and the conditional probability of future observations given current states, respectively, which are recombined to obtain hidden state posterior probabilities. For a $K$-rate

HMM, the overall computational cost of the resulting procedure is $O(T_K N^{K+1})$. On the other hand, collapsing multiple state components into a single state variable and invoking the HMM forward-backward algorithm directly leads to a $O(T_K N^{2K})$ algorithm, which is exponentially worse. See [3] for details.

### 3.3. Parameter Estimation

Parameters of the multi-rate HMM consist of the initial state and state transition distributions, and the state-conditional output distributions. In this work, we use mixtures of multivariate Gaussian densities to model output distributions.

Since states are never observed, maximum likelihood estimation of the multi-rate HMM parameters using training data $\mathcal{O}$ requires the optimization of the marginal log-likelihood function, $\log p(\mathcal{O})$, which is mathematically intractable. Instead, we use the expectation-maximization (EM) algorithm [4] to maximize it iteratively. The EM algorithm replaces the incomplete data likelihood function by the complete data one, $\log p(\mathcal{O}, \mathcal{S})$, averaged over all possible state completions $\mathcal{S}$, each of which is weighted by the a posteriori probability of the corresponding state completion. At the $n$-th iteration, parameters are updated according to

$$\theta^{(n+1)} = \underset{\theta}{\operatorname{argmax}}\, E_{\mathcal{S}|\mathcal{O}} \left[ \log p_\theta(\mathcal{S}, \mathcal{O}) | \mathcal{O}, \theta^{(n)} \right],$$

where $\theta^{(n)}$ denotes the parameter estimate at the $n$-th iteration. This maximization problem can be solved analytically to give explicit parameter updates for initial state and state transition distributions, and for mixture weights, means and covariances associated with the state-conditional output densities. See [3] for details.

## 4. EXPERIMENTS

### 4.1. Task

The task is the prediction of the amount of wear on $1/2''$ endmills milling titanium. The data is collected as follows. A sharp tool has been used to climb-cut notches in titanium blocks for multiple cutting passes until it becomes worn. An accelerometer mounted on the front plate of the spindle housing is used to record wide-band machining vibrations. At the end of selected cutting passes, a human operator measures the amount of wear on cutting edges. Quantizing these measurements, we label a cutter at the early stages of its lifetime as "A", a partially worn but still usable one is labeled as "B", and a dull one is labeled as "C". The assumption that wear level does not decrease over time is made to infer the state of wear on some unlabeled passes, which are also regarded as labeled in the experiments. The data set consists of 15 tools with a total of 145 cutting passes, of which 81 are labeled. Independent training and testing sets are constructed, and the training data set is used in three-way cross validation (CV) for development purposes.

### 4.2. Signal Processing

To capture transients and other short-term events, a sequence of cepstral feature vectors concurrent with flute rotations are derived (each cutting tool has four flutes). At every flute period, 20 cepstral coefficients are computed, based on a 25-th order linear prediction analysis with a $10\%$ window overlap [19]. Out of 20 coefficients, only the 1-st, 4-th, 5-th, and 6-th are kept, as they were found to have significantly higher discriminatory information based on the scatter ratio of features coming from sharp and worn passes. To minimize accelerometer placement effects [7], cepstral mean subtraction has been applied on a tool-by-tool basis similar to channel compensation in speech recognition. To capture long-term characteristics, a 40-frame average of the original cepstral features has been taken. Averaging windows are non-overlapping, and hence the averaged feature sequence is 40 times slower than the original one. We refer to the original cepstral features and their averages as short- and long-term features, respectively.

### 4.3. Training and Classification

The wear dynamics during a tool's lifetime was modeled both across multiple passing passes and within a cutting pass. The changes in cutter wear level from one cutting pass to another were modeled by a three-state HMM, where states represented wear levels and only transitions to the same or higher wear levels are allowed. Wear dynamics within a cutting pass were modeled either by HMMs or multi-rate HMMs characterizing the statistics of cepstral features under different amounts of wear. Ergodic state topologies were selected based on pilot experiments and the absence of prior knowledge suggesting otherwise. The optimal number of states and number of Gaussian mixtures per state were determined via CV. All system parameters were jointly estimated using the EM algorithm, which automatically deals with unlabeled cutting passes and hidden states in both HMMs and multi-rate HMMs.

During testing a cutting pass is classified according to the Bayes decision rule conditional on tool's whole cutting history. Previous work [8] found that a second stage classifier, generalized linear model (GLM), improves on the class a posteriori probability estimates and is also employed in this work. The GLM parameters are estimated from the training data via CV, using the iterative re-weighted least squares algorithm [12]. A GLM also provides a convenient framework for classifier combination taking log-odd ratios from multiple systems as input.

### 4.4. Results

We evaluated the 2-rate HMM using short- and long-term features against four HMM systems: HMM systems using short- and long-term features, an HMM system using concatenated short- and long-term features, and the classifier combination of the two HMM systems using short- and long-term features via a GLM. In the feature concatenation approach, long-term features are over-sampled to make them synchronous with the short-term ones.

In addition to 3-level classification accuracy, models are evaluated according to the normalized cross-entropy (NCE) metric,

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

for predicting class $C$ using features $X$, where $H(C|X)$ and $H(C)$ are the (cross-)entropy functions for the a posteriori and a priori class probabilities averaged with respect to the empirical test data distribution. NCE measures the information that is existent in features $X$ relevant for predicting $C$ as represented by the assumed model. Using the prior as a classifier results in an NCE of 0; 1 is a perfect score; and highly confident incorrect classification decisions result in negative NCE. NCE is a useful metric here in that it is more sensitive than accuracy given limited test data.

**Table 1**. Multi-level wear classification error rate (err) and NCE (nce) results for the HMM with short-term features (HMM-F), with long-term features (HMM-C), with concatenated long- and short-term features (HMM-CF), for the classifier combination of HMM-C and HMM-F, and for the 2-rate HMM (MHMM) systems on the test set.

| *Model* | *err %* | *nce* |
|---|---|---|
| HMM-F | 43 | -0.54 |
| HMM-C | 43 | -2.91 |
| HMM-CF | 43 | -2.69 |
| HMM-C+F | 45 | -6.85 |
| MHMM | 33 | 0.23 |

The multi-level classification error rates and NCE values for the HMM and multi-rate HMM classifiers in the test set are reported in Table 1. The a priori classifier assigning all cutting passes to the wear level having the highest frequency in the training data set has an error rate of $66\%$. All classification accuracies reported in Table 1 are statistically different from the a priori classifier at the significance level $p < 0.01$. The classification error rates of all HMM systems are similar, but the one using short-term features performs significantly better than the others in terms of NCE. Both the HMM with feature concatenation and the classifier combination approach fail to integrate information from short- and long-term features. The large negative NCE associated with the feature concatenation approach confirms our hypothesis that oversampling results in over-confident classifiers. The 2-rate HMM successfully uses the information available in the combination of short- and long-term features to improve on accuracy and confidence estimation. While the error rate differences are not statistically significant due to the small test set size ($p = 0.14$), the NCE differences are reasonably significant ($p = 0.06$), compared to using the HMM with short-term features.

Experiments with different averaging window sizes and feature extraction schemes for short- and long-term features give similar results.

## 5. SUMMARY

In this paper we introduce the multi-rate HMM for multi-scale modeling and apply it to the classification of titanium milling tool wear. The multi-rate HMM is a generalization of the standard HMM framework to multiple time scales, where variability is decomposed into scale-based components, and multiple state and observation sequences are used to represent process state and observational variability at different time scales. Results in a titanium milling tool-wear classification task show the utility of multi-rate HMMs for modeling wear dynamics.

**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. Initial experiments with the multi-rate HMM appeared in [7] for a different milling task, which involved a collaboration of the authors with Randall K. Fish and Gary D. Bernard. The system architecture used in this work is an improved version of that reported in [7]. The authors thank Dr. Fish for discussions about HMM-based tool-wear monitoring, and Dr. Bernard of Boeing for providing the data and for discussions about titanium milling.

## 6. REFERENCES

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

[2] M. Brand and N. Oliver. Coupled hidden Markov models for complex action recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recog.*, pp. 994–999, 1997.

[3] Ö. Çetin *et al.* Multi-rate hidden Markov models for monitoring of machining tool wear. Tech. report, UW EE, 2003.

[4] A. Dempster *et al.* Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B*, 39:185–197, 1977.

[5] H. Ertunc et. al Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs). *Intl. J. Machine Tools ad Manuf.*, 41:1363–1384, 2001.

[6] S. Fine *et al.* The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

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

[8] R. Fish *et al.* Multilevel classification of milling tool wear with confidence estimation. *IEEE Trans. PAMI*, 25(1):75–85, 2003.

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

[10] L. Heck and J. McClellan. Mechanical system monitoring using hidden Markov models. In *Proc. ICASSP*, pp. 1697–1700, 1991.

[11] J. Li *et al.* Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models. *IEEE Trans. IT*, 46(5):1826–1841, 2000.

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

[13] H. Nock and M. Ostendorf Parameter reduction schemes for loosely coupled HMMs. *Computer Speech and Language*, 17(4):233–262, 2003.

[14] M. Ostendorf *et al.* Joint use of dynamical classifiers and ambiguity plane features. In *Proc. ICASSP*, pp. 3589–3592, 2001.

[15] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[16] L. Owsley *et al.* Self-organizing feature maps and hidden Markov modes for machine-tool monitoring. *IEEE Trans. SP*, 45:2787–2798, 1997.

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

[18] L. Wang *et al.* Hidden Markov model-based tool wear monitoring in turning. *J. Manuf. Science and Engr.*, 124(3):651–658, 2002.

[19] S. Young *et al. The HTK Book (for HTK Version 3.2)*. Cambridge University, 2002.