JOINT USE OF DYNAMICAL CLASSIFIERS AND AMBIGUITY PLANE FEATURES

M. Ostendorf, L. Atlas, R. Fish, Ö. Çetin, S. Sukittanon, G. D. Bernard†

Electrical Engineering Department, University of Washington, Seattle, WA 98195
{mo,atlas,fishr,cozgur,ssukitta}@ee.washington.edu
†Manufacturing Research & Development, Boeing Commercial Airplane Group
gary.d.bernard@boeing.com

ABSTRACT

This paper argues for using ambiguity plane features within dynamic statistical models for classification problems. The relative contribution of the two model components are investigated in the context of acoustically monitoring cutter wear during milling of titanium, an application where it is known that standard static classification techniques work poorly. Experiments show that explicit modeling of long-term context via a hidden Markov model state improves performance, but mainly by using this to augment sparsely labeled training data. An additional performance gain is achieved by using the shorter-term context of ambiguity plane features.

1. INTRODUCTION

When used in classification or prediction of acoustic signals, time-frequency analysis is usually suitable for time scales up to the hundreds of milliseconds. That is, assuming typical airborne acoustic frequencies and short-time analysis frame rates, a time-frequency representation with longer than ten distinct frames in time can potentially overwhelm a classifier designed for non-parametric time-frequency representations. Compression of time-frequency redundancy and decimation in time can extend this time scale to several seconds. However, many applications, such as speech recognition or, as we will discuss here, machining monitoring, require a significantly longer temporal context.

Hidden Markov models (HMMs), which have been applied to a wide range of problems, are suitable for this role of modeling over thousands of consecutive frames in time. HMMs thus have the potential to model longer temporal context. However, as has been shown quite dramatically in speech recognition studies, HMMs do poorly unless some short-term dynamical processing is used to enhance nonstationary events. For speech recognition systems, the short-term dynamic processing is typically an approximate derivative acting across five consecutive frames. This short-term dynamic feature, which is commonly called a "delta," is centered at every time frame and simply augments an unprocessed frame. This fixed and unoptimized doubling of a feature size usually results in a substantial improvement of recognizer accuracy.

In this paper we observe how this delta feature is simply a fixed weighting in an auto-ambiguity plane. This weighting is uniform in lag and separable from transformed time (doppler). As we will show, a more specific choice of auto-ambiguity weighting [1, 2], when combined with a HMM, improves performance in our tests of tool wear monitoring for milling of titanium.

2. BACKGROUND

We use the term "dynamical classifiers" to include any model that characterizes the time-varying behavior of a process. The time-varying nature may be associated with a changing mode of the process (e.g. phonemes in speech, amount of wear in tool monitoring) and/or temporal dynamics within a particular mode. The best known dynamical classifier is the HMM, which represents a process as being generated by an unobserved discrete Markov chain. Typically, the observations are described by state-dependent Gaussian mixture distributions, but neural networks are also used in some systems. Due to the Markov assumptions, there are efficient algorithms for state recognition and parameter estimation for HMMs [3].

HMMs are quite powerful for representing duration variability of different modes and capturing long-distance contextual effects via the state sequence. However, the standard HMM framework is often criticized for its limited ability to capture local feature dynamics. To address this weakness, several new models have been proposed, referred to as "trajectory models" or "segment models", as surveyed in [4]. The use of approximate derivative features (or, deltas), standard in virtually all HMM systems, is a poor man's solution to incorporating local dynamics, but it has a big impact on performance given the low cost. We argue that further performance gains might be obtained by more principled selection of dynamic features using advanced time-frequency methods. Thus, the theme of this work is combining dynamic features (to capture local time-frequency characteristics) with dynamic classifiers (to capture temporal variability and more global context).

3. TIME-FREQUENCY FEATURES

In quadratic time-frequency research, it is sometimes desirable to utilize the auto-ambiguity (AA) plane when designing time-frequency representations (TFRs). The AA plane is given by

$$A[\eta,\tau] = \mathcal{F}_{n\rightarrow\eta}\{x^*[n]x[((n+\tau))_N]\},$$

where $\mathcal{F}$ is a Fourier transform. Since no doppler effects are appropriate for our applications of the ambiguity plane here, we call the time-transform variable η "variational frequency" [5]. Similarly, since range delay effects are not appropriate for our work here, the frequency transform variable τ is simply an instantaneous estimate of autocorrelation. $A[\eta,\tau]$ is the characteristic function of the discrete Rihaczek time-frequency representation, given by the two-dimensional Fourier transform

$$R[n,k] = \mathcal{F}_{\eta\rightarrow n}\{\mathcal{F}_{\tau\rightarrow k}\{A[\eta,\tau]\}.$$

A fundamental property of quadratic time-frequency analysis is that any quadratic TFR can be generated from $A[\eta,\tau]$ by application of the appropriate kernel function. The resultant generalized quadratic TFR is given by

$$G[n,k] = \mathcal{F}_{\eta\to n}\{\mathcal{F}_{\tau\to k}\{\phi[\eta,\tau]A[\eta,\tau]\},$$

where $G[n,k]$ is a smoothed version of the discrete Rihaczek time-frequency representation $R[n,k]$. A selection of features from the auto-ambiguity plane thus represents an implicit and specific choice of a smoothed Rihaczek TFR.

The use of delta features, as in speech recognition systems, represents only a fixed high-pass weighting in variational frequency η and has no dependence upon lag τ. Thinking of the delta feature as a dynamic feature, it is reasonable to consider other dynamic features, and the AA plane offers a convenient framework for doing so. However, much of the AA plane will not be useful, so some sort of feature reduction or selection is needed, particularly for training HMM parameters given limited data.

4. CASE STUDY: TOOL WEAR MONITORING

Over the past three decades, industry has realized the importance of selectively automating routine tasks of manufacturing operations. In machining parts, for example, the common industrial practice of replacing cutters according to a fixed schedule based on average cutter life is problematic and/or inefficient because of the wide variation in usable cutter life. For this reason, a substantial amount of research has gone into the field of automatic monitoring and control of machining processes [6, 7]. Research has focused on developing sensors, feature extraction methods, and automatic classification techniques for predicting when a tool is dull and needs to be replaced. Much of the work has involved static features and classifiers, which have been relatively effective for applications involving machining steel, but do not translate well to problems associated with machining titanium.

In machining of titanium alloys, after an end mill has cut for a while, the hot elemental titanium loves to diffusion-bond to the cutting edges. This process, of titanium from the workpiece welding to the cutter, forms a so-called "built-up edge" (BUE in milling jargon) that is carried by the primary cutting edge as it slices chips from the workpiece. As the BUE increases in volume over time, the forces experienced by the cutting edge also increase until the bonding forces are overcome and a large fraction of the BUE breaks away from the cutting edge. When the entire engaged length of all flutes of an end mill are involved in cycles of welding/release (of BUE), the time-frequency structure of both vibrations and very high frequency transients change considerably from those of the same cutter in the absence of BUE. Particles of the cutting edge substrate can also be torn away as the welded titanium breaks away, increasing tool wear. One cycle of build-up and release of BUE welded titanium may be as short as a second or as long as 30 seconds. Interspersed among these cycles of build-up and release are quiet, BUE-free periods characterized by reduced cutting forces, horsepower and vibration as well as reduced rate of cutter wear. These quiet periods may occur even when a cutter is rather worn, near the end of its useful life.

The behavior described above illustrates one substantial reason that conventional methods for tool wear monitoring fail when applied to milling of titanium. Another reason is that machinability of titanium work pieces can be quite variable and heterogeneous, containing random hard spots that may damage the cutter. Any successful strategy for tool wear monitoring of titanium milling must consider both history and context.

Until recently, only static classifiers have been used for tool wear applications [8, 9, 10]. Feature vectors representing an entire milling pass or drawn from some portion of a pass were collected and classification was posed as a binary problem of determining whether these features were generated by a cutter which was "dull" or "not dull". In reality, cutter wear is a dynamic process. Cutters move from being new to progressively greater levels of wear, and the feature vectors during each cutting event change as the cutter moves through the workpiece. Heck and McClellan [11] captured the progressive nature of drilling bit wear in a 5-state HMM, where the different states correspond to different levels of wear. Fish *et al.* [12] extended this idea, using states to model both the level of wear and the dynamics within a milling pass. HMMs can also be used to model dynamics at a finer time scale, i.e. the time-frequency structure of a transient [13, 14, 15]. The strategy in this work will be to use more sophisticated time-frequency analysis to model dynamics at the finer time scale, and to capture long distance effects with the HMM.

The past decade has seen a growing interest in applying advanced time-frequency analysis methods to machine monitoring. Zheng and Whitehouse [16] observed that the moments of the Wigner distribution of sensor outputs are useful for detecting incipient chatter and characterizing changes in the workpiece. Atlas *et al.* [7] summarize other results showing that more advanced time-frequency representations are required for determining salient features for classification. Gillespie and Atlas [1] introduce the use of the autoambiguity (AA) plane combined with feature selection to the problem of tool-wear monitoring. This paper will extend that work to HMM classifiers and will also investigate feature selection procedures.

5. EXPERIMENTS

5.1. Experimental Paradigm

The data used here was recorded from 1/2" end-mills milling titanium. At the end of a limited number of selected milling passes, each cutter was removed and its wear level microscopically measured by a master machinist and recorded before it was replaced and milling continued. The labels assigned based on these measurements are referred to as "known" labels. A cutter in the early stages of wear was labeled as "A", one which had exceeded the acceptable wear threshold was labeled "C", and those approaching the wear threshold but not yet ready to be replaced as "B".

The cutters were divided into two independent sets, one for training and the other held out for test. The training set, consisting of six 1/2" cutters, was used to train model parameters and evaluate different topologies. During this development phase, the training set was used in a three way cross validation to evaluate performance. Once development was complete, all six cutters in the training set were used to train the models used to classify the held-out test set which consisted of seven different 1/2" cutters. Using a default label of "not dull" for all data samples gives us "chance" performance, which is an accuracy of 85% (52/61) on the cross validation set and 83% (52/63) on the evaluation test set.

In the experiments described below, all classification systems implemented use HMMs with the same state topology for each wear level. (The static classifier is a special case of an HMM with only one state.) The best case classifier was then used in combination with a second stage classifier for evaluating the utility of different feature sets. The second stage classifier, in this case a generalized linear model [17], is used to improve the prediction of the posterior probability that the tool is dull, as proposed in [12]. The posterior probability is more useful to an operator than a hard decision and also provides a more fine-grained view of the classifier performance.

The posterior probability estimate is evaluated using normalized cross entropy (NCE),

$$NCE = \frac{H(D) - H(D|X)}{H(D)},$$

where D is the binary variable indicating whether or not the tool is dull, X is the observation sequence, and H is the entropy computed using the empirical test data distribution in the expectation. (The empirical distribution makes this a “cross” entropy.) The NCE measure indicates how much information is provided in the predicted posterior probability that the tool is dull relative to simply using the prior probability alone. The NCE provides an additional metric for predicting performance differences between feature set choices, which is particularly useful here due to the small test set sizes.

The different features explored include: baseline energy features, a subset of points in the autoambiguity plane automatically selected with and without the restriction to consider only stationary features and chosen to discriminate the different wear levels, and a set of autoambiguity points selected to be used more generally in both steel and titanium milling applications [1]. Except where noted, features were estimated at a rate of one per flute strike, or four times per revolution.

5.2. Topology Evaluation

In an earlier application of our system to the milling of steel, we found that the feature vectors recorded when the cutter first entered the workpiece *(entry)*, were different than those recorded when the tool leaves the workpiece *(exit)*, which were both different from those collected during the bulk of the milling pass, *(bulk)*. These different stages of a pass, *entry*/*bulk*/*exit*, were best modeled as a left-to-right HMM.

Inspection of the feature vectors for titanium suggested that milling of titanium might not have this same left-to-right behavior. To test this hypothesis, we evaluated three different HMM topologies. The first was the same as had been used for steel. This consisted of three left-to-right single mixture states followed by a single state with three mixtures, followed by another three left-to-right single mixture states. The second used the same number of free parameters but was a single state with nine mixtures. Finally, we also investigated a single state, four mixture model. In each of the three models tested, the feature vectors used were the “general” AA features. Topologies with a single state and multiple mixtures outperformed the topology intended to model a milling pass with recognizable left to right progression, so only this topology was used in subsequent experiments.

5.3. Using Context in Training and Test

Since tool wear is (for the most part) a gradual process, knowledge of the level of wear in a previous pass can reasonably be expected to improve the accuracy of classification of features from the present pass. In fact, viewing each milling pass in its context in the life of a cutter allows us to: i) add training labels because of our assumption of increasing cutter wear; ii) use unlabeled data to train our models using the Expectation-Maximization algorithm; and iii) allow the classification of previous milling passes to influence the classification of the present pass.

To investigate the impact of the use of context, we trained models using only those passes explicitly labeled by an expert machinist and classified each milling pass independently of all other passes of the same cutter. We also repeated classification using models trained with the additional data made possible by the context assumptions but without imposing context on classification. The results in table 1 show that using context in training is critical – without it the performance is worse than chance. Using context in classification gave a consistent but statistically insignificant gain. However, it may be that different modeling assumptions could lead to a classifier that is better able to take advantage of context.

Table 1. Performance (% correct) of three HMMs using a single state/nine mixture topology for each wear level, comparing different uses of context.

Use of Context	1/2” CV	1/2” Test
Training & Classifier	95	94
Training Only	93	92
No Context	77	75

5.4. Comparisons of Feature Sets

Once the topology was selected and the use of long-term context in both training and test had been established, we investigated various feature sets. Each feature set used the log of the total energy in the vibration signal as its first dimension. Our first feature set added only the delta coefficient to the log total energy. The remaining feature sets added features drawn from the auto-ambiguity plane.

The second and third AA feature sets selected a single coefficient (so as to keep the number of parameters comparable to the delta energy case) using a linear transformation of features from the AA plane. The transformation was estimated automatically using supervised linear discriminant analysis (LDA) and training only with titanium data. In order to capture phenomena at longer time scales, as is the case for the delta coefficient, the AA features were computed over a larger window, specifically 40 times that of the energy features. In the table, “21 AA Stationary” refers to the 21 AA features which are constrained to lie on the $\eta = 0$ axis and are thus stationary. “65 AA Full” refers to 65 AA features which consist of both stationary and non-stationary features. To reduce dimensionality for LDA design, we used a subset that corresponded to a triangle in the AA plane that included the lower half of the stationary features and the low-η values.

The fourth feature set (also in Table 1), referred to as “AA General”, includes AA features computed at the original data rate and selected based on inspection of clustered data on both steel and titanium data sets, which include both stationary ($\eta = 0$) and non-stationary ($\eta > 0$) elements. Feature selection uses automatic clustering combined with visual inspection of codewords to select salient points from the time-frequency plane [1]. First, vector quantizer design is used for unsupervised clustering of the ambiguity plane representations of 1/4 revolution windows into several codewords, and the quantizer is used to label all data samples. Then the relative frequency of occurrence of each codeword is computed as a function of time (or milling pass) for different sizes of cutters and for both titanium and steel materials. Observing the actual AA codewords showed that, as tool wear increased, the frequency of 1/4 revolutions with significant extent in variational frequency and lag increased from nothing to approximately 15%. The exciting finding was that this trait held for all materials and tool sizes analyzed. Based on this finding, six points from the AA plane (including energy) were chosen by hand for general use across milling applications.

The performance of these various feature sets are shown in table 2, including both accuracy and NCE performance statistics. As expected, we find that removing constraints on features chosen from the AA plane improves performance (rows 2 vs. 3), although the LDA weight vector does put much more weight on the subset of stationary features. In addition, AA dynamic features outperform the delta feature (rows 1 vs. 3) when the number of features are constrained to be the same.

Table 2. Performance of four different feature sets on the test set using a single state/four mixture classifier for each wear level. 'LogE' = log energy, 'Delta' = a derivative estimate, 'LDA' = linear discriminant analysis, and 'AA' = auto-ambiguity features.

Features	%	NCE
logE + 1 Delta logE	90	0.12
logE + 1 LDA(21 AA Stationary)	87	0.09
logE + 1 LDA(65 AA Full)	94	0.10
logE + 5 AA General	94	0.21

In the last row, where the feature dimensionality is increased, the gain is not measurable in terms of accuracy, but there is a real improvement in performance as indicated by the NCE of the confidence scores. Since the "AA general" features are computed with a different time window, it may be possible to further improve performance by combining the different time scales.

6. CONCLUSIONS

In summary, this paper argues for use of dynamic time-frequency features within dynamic statistical models for classification problems. The standard use of delta features within HMMs is a simple example that can be improved upon by less constrained selection of features from the autoambiguity plane. Experiments on acoustically monitoring cutter wear during milling of titanium show that both the use of dynamic features and dynamic classifiers can improve performance, though the use in training is critical. In particular, the representation of dynamics in the classifier is important for using unlabeled training data, and the best results for dynamic time-frequency features were obtained by choosing features that generalize over different training conditions.

7. ACKNOWLEDGMENTS

We wish to thank the Office of Naval Research for their support of this project under the Center for Auditory and Acoustics Research (ONR-2883401) and the Boeing Commercial Airplane Group for sharing their machining expertise and milling data set.

8. REFERENCES

[1] B. Gillespie and L. Atlas, "Data-driven time-frequency classification techniques applied to tool-wear monitoring," *Proc. Int. Conf. Acoust., Speech & Signal Proc.*, 649–652, 2000.

[2] B. Gillespie and L. Atlas, "Optimizing Time-Frequency Kernels for Classification," *IEEE Trans. Signal Proc.*, March, 2001, in press.

[3] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, **77**(2) 257–285, 1989.

[4] M. Ostendorf, V. Digalakis and O. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech & Audio Proc.*, **4**(5) 360–378, 1996.

[5] T. Kailath, "Channel characterization: time-variant dispersive channels," in *Lectures in Communication Systems Theory*, E. Baghdady (Ed.), McGraw-Hill, New York, 1960.

[6] A. Ulsoy and Y. Koren, "Control of machining process," *Journal of Dynamic Systems, Measurement and Control*, **115** 301–308, 1993.

[7] L. Atlas, G. D. Bernard and S. B. Narayanan, "Applications of time-frequency analysis to manufacturing sensor signals," *Proc. of the IEEE*, **84**(9) 1319–1329, 1996.

[8] R. Du, M.A. Elbestawi, and S.M. Wu, "Automated monitoring of manufacturing processes, part 2: Applications," *ASME Journal of Manufacturing Science and Engineering*, **117** 133–141, 1995.

[9] S. Rangwala and D. Dornfeld, "Sensor integration using neural networks for intelligent tool condition monitoring," *ASME Journal of Manufacturing Science and Engineering*, **112** 219–228, 1990.

[10] E. Emel and E. Kannatey-Asibu, Jr., "Tool failure monitoring in turning by pattern recognition analysis of AE signals," *ASME Journal of Engineering for Industry*, **110** 137–145, 1988.

[11] L.P. Heck and J.H. McClellan, "Mechanical system monitoring using hidden Markov models," *Proc. Int. Conf. Acoust., Speech & Signal Proc.*, 1991, 1697–1700.

[12] R. Fish, M. Ostendorf, G. Bernard, D. Castanon and H. Shivakumar, "Modeling the progressive nature of milling tool wear," *Proc. of the ASME, Manufacturing Engineering Division*, **11** 111-117, 2000.

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

[14] J. McLaughlin, L. Owsley, L.E. Atlas, and G.D. Bernard, "Advances in real-time monitoring of acoustic emissions," *Proc. of the SAE Aerospace Meeting*, 1997.

[15] L. Atlas, M. Ostendorf, and G. Bernard, "Hidden Markov models for monitoring machining tool-wear," *Proc. of ICASSP*, 3887–3890, 2000.

[16] K. Zheng and D. J. Whitehouse, "The application of the Wigner distribution to machine tool monitoring," *Proc. Inst. Mech. Engrs.*, **206** 249–264, 1992.

[17] J. M. Chambers and T. J. Hastie, *Statistical Models in S*, Wadsworth & Brooks, 1992.