

# IMPROVING CLASSIFICATION RESULTS IN MUSIC INFORMATION RETRIEVAL

*Kris West*

School of Computing Sciences  
University of East Anglia  
kristopher.west@uea.ac.uk

## 1. MOTIVATION

As personal computing power increases, so does both the demand for and the feasibility of automatic music analysis systems. Soon content discovery and indexing applications will require the ability to automatically analyse, classify and index musical audio, according to perceptual characteristics such as genre or mood. Recently this has been a topic of interest for many research groups worldwide; such as Queen Mary's University London, which Prof. Stephen Cox and I recently visited for a one-day Digital Music Research Network event and subsequently invited Prof. Stephen Cox to lecture the Digital Music research group on Hidden Markov Models. The Digital Music research group at Queen Mary's University currently holds three EPSRC grants in this field. Research in this field is supported by several conferences, including: the International Symposium on Music Information Retrieval (ISMIR), the International Conference on Digital Audio Effects (DAFx), the International workshop on content-based multimedia indexing (CBMI) and the International Computer Music Conference (ICMC).

## 2. BACKGROUND STUDY

In the first year of my post-graduate course I have investigated a series of factors that affect the automatic classification of audio signals. Specifically I have conducted an empirical evaluation of factors affecting classification performance of automatic music classification techniques including: feature set (parameterisation of audio signal), modelling temporal variation of spectral features, transformations and dimensionality reduction, classification scheme and classifier topology.

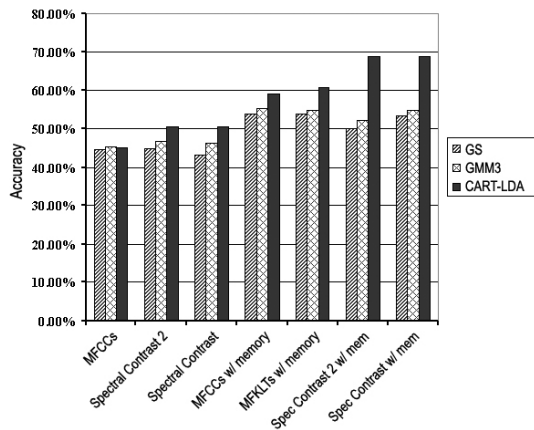
I have evaluated the classification performance of two different measures of spectral shape used to parameterise the audio signals prior to classification, Mel-frequency Cepstral Coefficients and

Spectral Contrast features. Mel-frequency Cepstral Coefficients (MFCCs) are perceptually motivated features originally developed for the classification of speech. MFCCs have been used for the automatic classification of musical audio signals in [TC02] and [SS02]. In [JLZ<sup>+</sup>02] a set of Octave-scale based Spectral Contrast features are proposed, which are designed to provide better discrimination among musical genres than Mel-Frequency Cepstral Coefficients.

In the field of automatic genre classification of audio signals, many different classification strategies have been employed, including multivariate single Gaussian models, Gaussian mixture models, self-organising maps, neural networks, k-means clustering, k-nearest neighbour schemes, supervised hierarchical implementations of the aforementioned classifiers and Hidden Markov Models. In several cases varying the specific classifier used did not affect the classification accuracy, however, varying the feature sets used for classification had a far more pronounced effect[MB]. In my evaluation I found that although both the use of Spectral Contrast features and Gaussian Mixture models tended to increase accuracy by 1 to 2% from a baseline of accuracy of  $43\% \pm 1\%$  (single Gaussian models trained on raw MFCCs), modelling of temporal variation in the calculated features, by converting them to windowed (2 second) means and variances of the features, tended to increase accuracy by 8%.

## 3. WORK COMPLETED IN FIRST YEAR

I have developed new classifiers based on the unsupervised construction of a binary decision tree classifier, as described in [CAR84], and linear discriminant analysis [WEB02] at each node of the tree, which has achieved an additional accuracy increase of as much as 6% on the means and variances of MFCCs and as much as 14% when used on the means and variances of Spectral Contrast features. Therefore, the highest average classifi-



**Figure 1.** Classification performance

cation accuracy achieved in the evaluation is  $68\% \pm 1\%$  which, when compared to the baseline of  $43\% \pm 1\%$  (Single Gaussian models trained on raw MFCCs) and the highest accuracy achieved  $54\% \pm 2\%$  (3 component Gaussian mixture models trained on the means and variances of MFCCs or Spectral Contrast feature), is a very significant improvement. The unsupervised construction of a very large ( $> 5000$  leaf nodes) decision trees for the classification of features calculated from musical audio signals is a new approach, which allows the classifier to learn and identify closely defined groups of sounds that only occur in certain types of music. I believe that the results achieved by these classifiers, represent the most significant increase in the classification accuracy of musical audio signals to date.

The results of this evaluation and the classifiers to be introduced are to be submitted as a paper to ISMIR 2004 on May 7th. The classifiers have also been used as part of an investigation in to the automatic classification of Midi tracks, being conducted by Ming Lee, Ronan Sleep and I, which will also be submitted to ISMIR 2004 on May 7th.

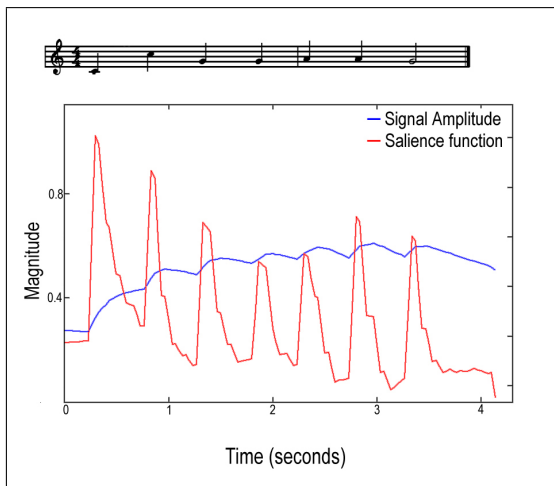
In the first year of my post-graduate course I have also evaluated several techniques for the identification of note onset times in musical audio streams and I have defined a new note onset detection technique, called Centroid shift onset detection, which is appears to be highly successful.

#### 4. IMPROVEMENTS EXPECTED FROM FURTHER RESEARCH

The accuracies listed above correspond to the frame by frame classification rate of the audio signal, i.e. they represent the proportion of audio frames correctly classified. In a real world classification system a large number of audio frames would be collected and then classified. Final classification of the whole sample is decided by the majority classification of the frames calculated from it, this is known as a “bag of frames” classification. It is my proposal to investigate three methods of increasing the accuracy of these real world classifiers, firstly by selecting frames for classification, secondly by directing the segmentation and parameterisation of the audio signal through onset detection and rhythmic analysis of the signal and thirdly by performing rhythmic analysis of the audio signal and adding categorical variable splits to the classification trees based on the rhythmic structure of the music.

Because the new hierarchical classifier appears to work by learning and identifying closely defined groups of sounds that only occur in certain types of music, it should be possible to increase classification accuracy by identifying and ignoring audio frames which do not contribute to the classification accuracy, such as near or total silence and frames common to many classes of audio. Initially a normal classification tree will be grown and evaluated. As part of the evaluation, the evaluation audio frames will be separated into two classes, those correctly classified and those incorrectly classified. A second, two class, classification tree will be grown from this data and used to filter both the training data and any test data to be classified. A new classification tree will be grown on the filtered data and saved as the final classifier.

The second approach to increasing the accuracy of audio classification involves performing rhythmic analysis of the audio signal. At present most music classification procedures ignore rhythmic information in music, classifying only on spectral or short time features, or at best include a naive representation of the beat structure, such as the period of the first two peaks of the autocorrelation function of the signal. I propose that by directing segmentation and parameterisation of a signal, according to rhythmic hypotheses, and then selecting those segments which contribute to classification accuracy, we can improve the classifier performance. Initially I will implement and evaluate three different methods of performing this segmentation and frame selection.



**Figure 2.** An example of a saliency function.

The first is based on the principle that musical audio events have a characteristic known as saliency, which means to have a quality that thrusts itself into attention. We can identify musical events in the audio stream by constructing a stable estimate of the audio signals saliency and selecting peaks in the saliency function. The audio signal can then be segmented accordingly and the most salient or attention grabbing frames selected for classification.

The second method involves the application of a single large Ergodic Hidden Markov model to the event onset and saliency sequence. The model will be used to produce a sequence of state numbers in parallel with the normal parameterisation of the audio signal. By estimating the mutual information [MUT04] between these state numbers and each class of the audio signals in the training data, we can identify states which may contribute little to the classification performance. A state which returns a low mutual information score for all classes is likely to be independent of the class of the audio signal and therefore the parameterised audio frames corresponding to that state should not be included in a “bag of frames” classification.

The final method is to train a single large Ergodic Hidden Markov model on the audio parameterised by the calculation of Spectral Contrast features, which I have already successfully applied to the classification of music. The state sequence output by this model will be used as before.

It will be necessary to evaluate whether audio frame selection is more successful when directed by raw note onset detection, or by a full rhythmic hypothesis, generated by the Hidden Markov

model. This approach will also require the definition of a stable process for estimating the saliency of an audio signal, frame by frame, and will build upon both work and experience I have gained in note onset detection techniques, in both the first year of my postgraduate course and as part of my undergraduate dissertation. This approach will be implemented as part of the feature extraction process.

The final approach to improving the “bag of frames” classifiers also involves performing rhythmic analysis of either the onset sequence detected by the saliency estimation or the parameterised audio signal with a single large Ergodic Hidden Markov model. I propose to add the capability to test, evaluate and implement categorical variable based splits to my Classification and Regression tree classifier series, with categorical variables based on either specific state or state sequence output by the Hidden Markov model or existing techniques of tempo, meter and drum pattern estimation, in addition to the normal splitting process. This approach is only applicable to decision tree based classifiers. Rhythmic information has already been successfully used to classify music as part of an expert system in [DPW].

## 5. REFERENCES

- [CAR84] *Classification and Regression Trees*. Wadsworth and Brooks/Cole Advanced books and Software, 1984.
- [DPW] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR) 2003*.
- [JLZ<sup>+</sup>02] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. Technical report, Department of Computer Science and Technology, Tsinghua University, China and Microsoft Research, Asia, 2002.
- [MB] Martin F McKinney and Jeroen Breebaart. Features for audio and music classification. In *Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR) 2003*.
- [MUT04] *A description of mutual information between two random variables with examples*. Connexions, Rice University, <http://cnx.rice.edu/content/m10178/latest/>, 2004.
- [SS02] Alan P Schmidt and Trevor K M Stone. Music classification and identification system. Technical report, Department of Computer Science, University of Colorado, Boulder, 2002.
- [TC02] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE transactions on speech and audio processing*, 2002.
- [WEB02] *Statistical Pattern Recognition*. John Wiley and Sons, Ltd, 2002.

**Table 1.** Project plan and timescale

Time allocated	Task
1 month	Modify decision tree classifier series to perform real world, “bag of frames” classifications.
$\frac{1}{2}$ month	Evaluate classification performance of simple “bag of frames” classifier.
2 months	Modify decision tree classifier’s training and evaluation procedures to identify misclassified frames in evaluation data set, to grow a two class decision tree classifier based on misclassified data partition and then use this tree to filter training data set and retrain main classification tree.
$\frac{1}{2}$ month	Evaluate classification performance of filtered decision tree classifier.
3 months	Complete the development of Octave-scale centroid shift salience estimation, evaluate performance of amplitude slope, FFT phase deviation and centroid shift salience estimators against synthesised Midi tracks and investigate a combination of these procedures for accurate salience estimation.
6 months	Implement two large, Ergodic Hidden Markov models. One will be trained on the salience estimates and the other on Spectral Contrast features, calculated from the audio signal, in parallel with the normal parameterisation of the audio signal.
2 months	Build a mutual information estimator, to evaluate mutual information between output states and audio classes.
2 months	Modify decision tree classifier series to accept frame filters based on the output of the mutual information estimators.
2 months	Modify decision tree classifier series to accept categorical variable splits, based on the state sequences output by the hidden markov models.
2 months	Evaluate classification performance of decision tree classifier with rhythmically based splits and rhythmically filtered decision tree classifiers.
22 months	Total