MIREX AUDIO GENRE CLASSIFICATION

Kris West

School of Computing Sciences
University of East Anglia
Norwich, UK
kw@cmp.uea.ac.uk

ABSTRACT

This extended abstract details a submission to the Music Information Retrieval Evaluation eXchange in the Audio Genre classification task. This submission is very similar to the system that placed second in the 2004 ISMIR Audio description contest. A novel feature set and segmentation of features is introduced and modifications to the Decision Tree based model used in the 2004 submission are detailed. Finally, the results achieved in the evaluation are analysed.

Keywords: MIREX, Audio, Genre.

1 FEATURE SET

Two feature sets are calculated in this submission, one describing the timbre of the audio and another describing the rhythmic content. 22 kHz audio is taken as input, divided into 50% overlapping 23 ms frames and a FFT performed to obtain the spectrum, prior to both analyses.

1.1 Spectral features

A novel feature called Mel-band Frequency Domain Spectral Irregularity is calculated to describe the timbre of the audio. This feature is calculated from the output of a Mel-frequency scale filter bank and is composed of two sets of coefficients, half describing the spectrum and half describing the irregularity of the spectrum. The spectral features are the same as those used Mel-frequency Cepstral Coefficients (MFCCs) without the Discrete Cosine Transform (DCT).

The irregularity coefficients are similar to Octave-scale Spectral Irregularity Feature as described by Jiang et al. (2002), as they include a measure of how different the signal is from white noise in each band. This allows us to differentiate frames from pitched and noisy signals that may have the same spectrum, such as string instruments and drums. Our contention is that this measure comprises important psychoacoustic information which can provide better audio modelling than Mel-frequency Cepstral Coefficients. This feature is calculated by estimating the difference between the white noise signal that would have produced the spectral coefficient, in each band, and the actual signal that produced it. Higher values of these coefficients indicate that the energy was highly localised in the band and therefore would have sounded more pitched than noisy.

These features are calculated with 16 filters to reduce the overall number of coefficients. We have also experimented with using more filters and a Principal Components Analysis, PCA, or a Discrete Cosine transform (of each set of coefficients), DCT, to reduce the size of the feature set, but have found performance to be similar using less filters, which also reduces computational complexity. This property may not be true in all models as both the PCA and DCT reduce covariance between dimensions of the features as do the transformations used in our model (see section 3), reducing or eliminating this benefit from the PCA/DCT.

An overview of the Spectral Irregularity calculation is given in figure 1.

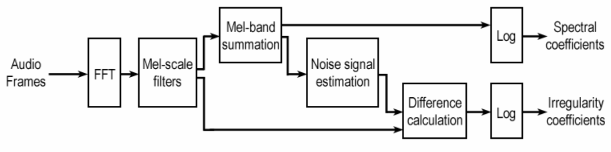


Figure 1. Spectral Irregularity calculation

1.2 Rhythmic features

In order to describe rhythmic features of the audio modulations of an onset detection function are calculated. The onset detection function peaks at the beginning of each audio event and so by estimating the modulations of this function we may be able to describe the periodicities in the audio. A combined Mel-band Phase and Amplitude onset detection function is calculated, as described in West and Cox (2005), divided into 7 second frames and the modulations estimated with an FFT. The first 30 coefficients are retained, representing modulations from 0 Hz to ~9 Hz.

2 SEGMENTATION

As shown in West and Cox (2005), we found that more information was available for audio modeling in short segments of the audio than in features averaged over the whole file and that features that had been segmented into individual events were both easier to model and caused less data load than a sliding harmonic modeling window.

Our final parameterisation of the audio is formed by segmenting the sequence of frames using a combined Mel-band Phase and Amplitude onset detection function and calculating the mean and variance of the features over the segment. The segment length is also appended. Because the rhythmic features vary much more slowly than the spectral features only the means of these are retained.

3 MODEL

A very similar model to our submission to the ISMIR 2004 Audio Description Contest is used based on a modified Classification and Regression Tree trees, as described by Breiman et al. (1984). This model has been modified, as described in West and Cox (2005), replacing the normal single variable splits with single Gaussian distributions and Mahalanobis distance measurements. The classes of data are divided into two groups and two Gaussian distributions are calculated to implement a binary split.

This model has been further improved transforming the features, at each node, using a multi-class Fisher's Criterion Linear Discriminant Analysis (LDA), yielding $N-1$ components, where N is the number of classes. This transformation allows the model to selectively weight features to provide the optimal separation between classes, massively reduces the data that the Gaussian distributions are calculated over and eliminates any off diagonal covariance. Transforming the data reduces the both the size of the final decision tree and computation time and increases the accuracy of the model.

3.1 Classifying examples with multiple vectors of features

Because we are using segmented features, each example has multiple vectors of features to consider. In order to classify each example the log of the classification likelihoods of each frame are summed across the piece, this is equivalent to multiplying the probabilities together. In order to avoid classification likelihoods of zero from eliminating the likelihood of a class in the summation, these likelihoods are smoothed using Lidstone's law as described by Lidstone (1920).

3.2 Handling variable length transcriptions

Because the transcriptions are of variable length and variable numbers of examples of each class are input to the model for training, the prior probabilities are highly skewed and can adversely affect the performance of the model. To alleviate this classification likelihoods output by the model are normalised by the prior probability of each class before the output class is selected. This drastically increases the classification accuracy of the model normalised by the class sizes and also increase the raw classification accuracy.

4 IMPLEMENTATION

This submission is entirely built in D2K using the M2K Toolkit with a few additional custom modules. Several of the modules used in this submission were contributed to the M2K Toolkit. The feature extraction and modeling itineraries are shown in figures 2 and 3.

5 RESULTS

The results achieved by this model in the MIREX 2005 genre classification evaluation are detailed in Table 1.

Evaluation metric	**Value**
Overall performance	75.29%
Magnatune Hierarchical Classification Accuracy	71.67%
Magnatune Normalized Hierarchical Classification Accuracy	68.33%
Magnatune Raw Classification Accuracy	68.43%
Magnatune Normalized Raw Classification Accuracy	63.87%
Magnatune Runtime (s)	43,327 secs
USPOP Raw Classification Accuracy	78.90%
USPOP Normalized Raw Classification Accuracy	75.45%
USPOP Runtime (s)	18,557 secs

Table 1. Results achieved in MIREX 2005 genre classification evaluation

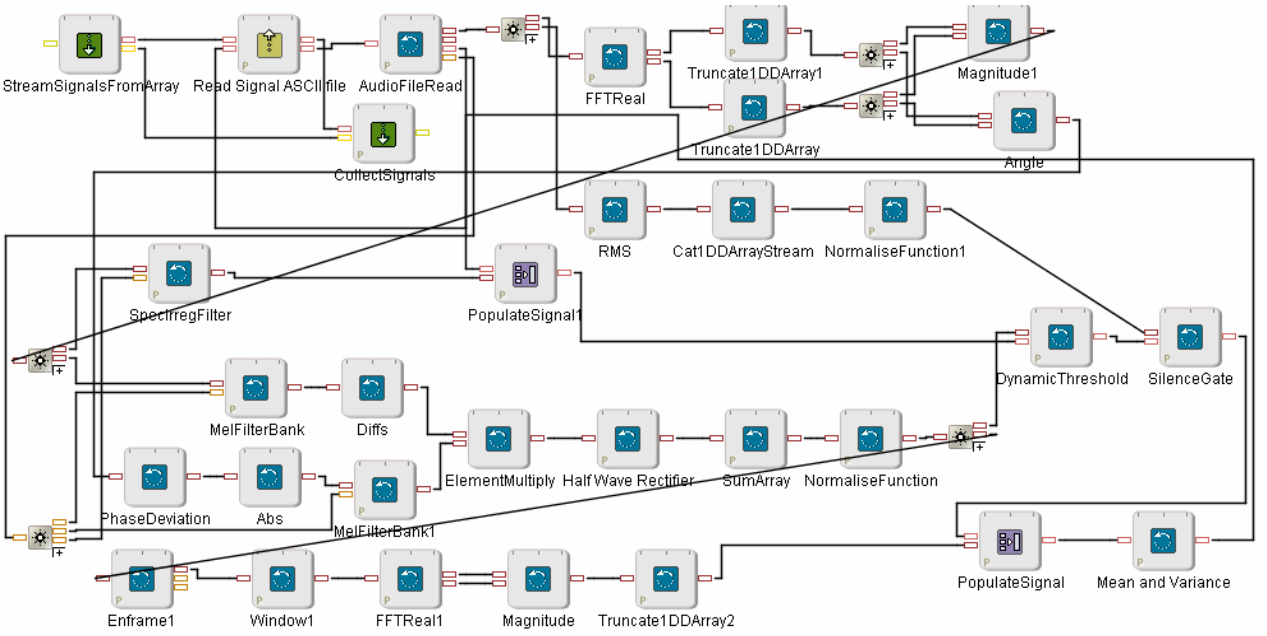
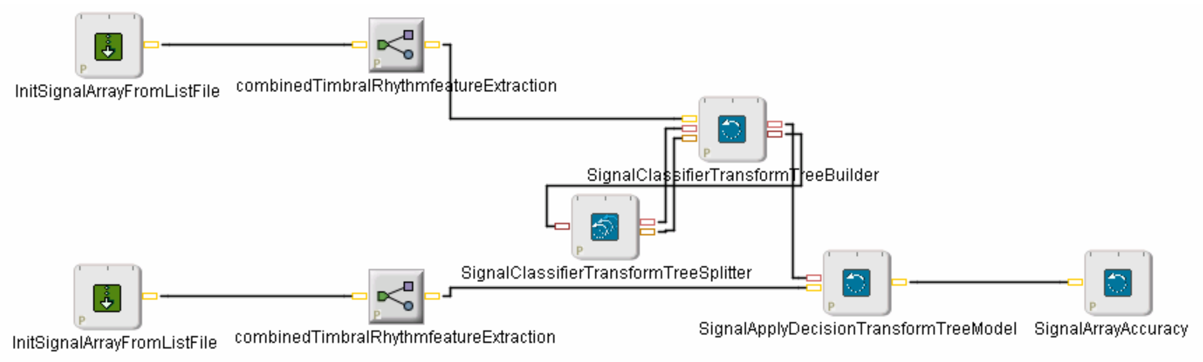


Figure 2. Feature Extraction itinerary

Figure 3. Modelling itinerary


InitSignalArrayFromListFile
combinedTimbralRhythmfeatureExtraction
SignalClassifierTransformTreeBuilder
SignalClassifierTransformTreeSplitter
InitSignalArrayFromListFile
combinedTimbralRhythmfeatureExtraction
SignalApplyDecisionTransformTreeModel
SignalArrayAccuracy

6 ANALYSIS

We achieved a relatively good performance (placing 3rd overall among 14 evaluated algorithms) although it should be noted that the results were very closely grouped, so the forthcoming significance tests may be very interesting. The runtimes of this technique were much higher than expected based on early tests. The majority of the runtimes reported in the results are for the feature extraction and model training process. The feature extraction process could be significantly sped up by selecting a small sample from each piece (30 seconds to 1 minute) to perform feature extraction and modelling on, as many techniques in the contest did. We chose to model the whole song, which in some cases (on the Magnatune database) was as long as 20 minutes.

We could also significantly reduce modeling time by using stronger stopping criteria to terminate tree growth. At present tree growths terminates if unable to improve resubstitution errors, a pure node is reached or if the minimum number of examples to train the LDA and Gaussian distributions is not available. This, perhaps inappropriately, grows very large trees as it rare to be able to establish pure nodes and in some cases very small leaf nodes may be produced. A stopping criteria based on the proportion of data at a node (rather than the absolute amount of data) and elimination/prevention of very small nodes may significantly reduce modelling time and increase performance.

Once trained, the model is extremely quick to apply to the data. This is an important observation as model training can be regarded as an indexing step, feature extraction can be massively parallelised and distributed (even across the internet), while model application to new data can be regarded as the true application cost (which requires relatively little memory or computation to perform).

Finally, it should be noted that no artist filter was used in the division of data between the test and training sets in this evaluation (i.e. ensuring an artist is only in either the test or training set, not both). Elias Pampalk has pointed out that this leads to very different behaviour in many models as classification may be more easily assigned by matching an artist to an example of their own work in the training set. This can only be examined if both filtered and un-filtered data is used in parallel evaluations and unfortunately there was not time to do this, particularly as you may wish to perform multiple iterations of this experiment to be sure of results.

7 FURTHER WORK

Further work will concentrate on further improvements to the model, such as using fuzzy tree traversals to improve classification accuracy, better splitting criteria (Entropy rather than Gini Index), better stopping criteria to reduce training time, modified pruning (to remove particularly small nodes from the tree) and sequential modelling of tree leaf nodes to leverage melodic information in each class of signal. Prototype selection could also be used to reduce modelling training time, while a NN based classification could be performed on the massively decreased dataset at each leaf node.

Further work on features will concentrate on decorrelating and reducing the dimensionality of the Timbral feature set and determining whether the rhythmic feature set is better modelled on its own with likelihoods combined in a classifier ensemble.

Further work on segmentation will focus on improving and properly evaluating the onset detector and examining the effect of silence gating on the model size. A reduction in false positives and more accurate segmentation may decrease the model complexity and improve overall accuracy.

ACKNOWLEDGEMENTS

Many thanks to IMIRSEL for the huge amount of effort spent in running the evaluation and for providing the M2K Toolkit. Thanks also go to the NCSA for providing D2K, in which the M2K Toolkit is implemented.

REFERENCES

D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. “Music type classification by spectral contrast feature.” In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME02), Lausanne Switzerland*, Aug 2002.

K. West and S. Cox. “Finding an optimal segmentation for audio genre classification”, *In Proceedings of the Sixth International Symposium on Music Information Retrieval, London, UK,* 2005 (to appear).

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. “Classification and Regression Trees”. *Wadsworth and Brooks/Cole Advanced books and Software*, 1984.

G. J. Lidstone. “Note on the general case of the bayeslaplace formula for inductive or a posteriori probabilities.” *Transactions of the Faculty of Actuaries, 8:182–192,* 1920.