FEATURES AND CLASSIFIERS FOR THE AUTOMATIC CLASSIFICATION OF MUSICAL AUDIO SIGNALS

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

Stephen Cox
School of Computing Sciences
University of East Anglia,
sjc@cmp.uea.ac.uk

ABSTRACT

Several factors affecting the automatic classification of musical audio signals are examined. Classification is performed on short audio frames and results are reported as "bag of frames" accuracies, where the audio is segmented into 23ms analysis frames and a majority vote is taken to decide the final classification. The effect of different parameterisations of the audio signal is examined. The effect of the inclusion of information on the temporal variation of these features is examined and finally, the performance of several different classifiers trained on the data is compared. A new classifier is introduced, based on the unsupervised construction of decision trees and either linear discriminant analysis or a pair of single Gaussian classifiers. The classification results show that the topology of the new classifier gives it a significant advantage over other classifiers, by allowing the classifier to model much more complex distributions within the data than Gaussian schemes do.

1. INTRODUCTION

As personal computing power increases, so do 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.

In the field of automatic genre classification of musical audio signals, classification is often performed on spectral features that have been averaged over a large number of audio frames. Many different classification strategies have been employed, including multivariate single Gaussian models [1], Gaussian mixture models [2], self-organising maps [3], neural networks [4], support vector machines [5], k-means clustering, k-nearest neighbour schemes [1], Hidden Markov Models [6] and supervised hierarchical implementations of the aforementioned classifiers. It has been observed that 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 on the classification accuracy[7].

In this paper, classification is performed on a large number of short audio frames calculated from a sample with the final classification being decided by a majority vote. We explore the features calculated from the audio signal, the temporal modelling of those features, and the classifying schemes that have been trained on the resulting data. Classification results are reported as "bag of frames" accuracies, where the audio is segmented into 23ms analysis frames and a majority vote is taken to decide the final classification. Finally we introduce new classifiers based on the un-supervised construction of a binary decision tree, as described in [8], and either linear discriminant analysis or a pair of single Gaussians [9] at each node of the tree. The un-supervised construction of a very large (> 5000 leaf nodes) decision trees for the classification of frames, from musical audio signals, is a new approach, which allows the classifier to learn and identify diverse groups of sounds that only occur in certain types of music. The results achieved by these classifiers represent a significant increase in the classification accuracy of musical audio signals.

In section 3 we describe the evaluation of different parameterisations of the audio signals and the transformations used on them. In section 4 the classifiers trained on this data are detailed and two new classifiers are introduced. In section 5 the test data used in the evaluation experiments is described, results achieved are discussed. In the final sections we detail the conclusions we have drawn from these results and detail potential areas for further research.

2. IMPLEMENTATION

All of the experiments detailed in this paper were implemented within the Marsyas-0.1 framework [10].

3. PARAMETERISATION OF AUDIO SIGNALS

Prior to classification the audio must be segmented and parameterised. We have evaluated the classification performance of two different measures of spectral shape used to parameterise the audio signals, Mel-frequency filters

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2004 Universitat Pompeu Fabra.

(used to produce Mel-Frequency Cepstral Coefficients or MFCCs) and Spectral Contrast feature. For comparison the *Genre* feature extractor for Marsyas-0.1 [10], which calculates a single feature vector per piece, is also included in this evaluation.

3.1. Segmentation

Audio is sampled at 22050 Hz and the channels averaged to produce a monaural signal. Each analysis frame is composed of 512 individual audio frames, with no overlap, representing approximately 23 ms of audio. Therefore the lowest frequency that can be represented in an analysis frame is approximately 45 Hz which is close to the lower threshold of human pitch perception. No overlap is used as additional experiments have shown no gain in accuracy for a 50% overlap, despite doubling the data processing load, in an already data intensive task.

3.2. Mel-Frequency filters and Cepstral Coefficients

Mel-frequency Cepstral Coefficients (MFCCs) are perceptually motivated features originally developed for the classification of speech [11]. MFCCs have been used for the classification of music in [1], [12] and [10]. MFCCs are calculated by taking the outputs of up to 40 overlapping triangular filters, placed according to the Mel frequency scale, in a manner which is intended to approximately duplicate the human perception of sound through the cochlea. The magnitude of the fast Fourier transform is calculated for the filtered signal and the spectra summed for each filter, so that a single value is output. This duplicates the output of the cochlea which is known to integrate the power of spectra within critical bands, allowing us to perceive a course estimate of spectral envelope shape. The Log of these values is then taken, as it is known that perception of spectral power is based on a Log scale. These values form the final parameterisation of the signal but must be transformed by the Discrete Cosine transform [13], in order to eliminate covariance between dimensions in order to produce Mel-Frequency Cepstral Coefficients.

3.3. Octave-scale Spectral Contrast feature

In [14] an Octave-based Spectral Contrast feature is proposed, which is designed to provide better discrimination among musical genres than MFCCs. When calculating spectral envelopes, spectra in each sub-band are averaged. Therefore only information about the average spectral characteristics can be gained. However, there is no representation of relative spectral characteristics in each sub-band, which [14] suggests is more important for the discrimination of different types of music.

In order to provide a better music representation than MFCCs, Octave-based Spectral Contrast Feature considers the strength of spectral peaks and valleys in each sub-band separately, so that both relative spectral characteristics, in the sub-band, and the distribution of harmonic and non-harmonic components are encoded in the feature. In most music the strong spectral peaks tend to correspond with harmonic components, whilst non-harmonic components (stochastic noise sounds) often appear in spectral valleys [14], which reflects the dominance of pitched sounds in western music. Whilst it is considered that two spectra that have different spectral distributions may have similar average spectral characteristics, it should be obvious that average spectral distributions are insufficient to differentiate between the spectral characteristics of these signals, which can be highly important to the perception of music.

The procedure for calculating the Spectral Contrast feature is very similar to the process used to calculate MFCCs. First an FFT of the signal is performed to obtain the spectrum. The spectral content of the signal is then divided into a small number of sub-bands by Octave scale filters, as apposed to the Mel scale filters used to calculate MFCCs. In the calculation of MFCCs, the next stage is to sum the FFT amplitudes in the sub-band, whereas in the calculation of spectral contrast, the spectra are sorted into descending order of strength and then the strength of the spectra representing both the spectral peaks and valleys of the sub-band signal are recorded. In order to ensure the stability of the feature, spectral peaks and valleys are estimated by the average of a small neighbourhood (given by α) around the maximum and minimum of the sub-band. Finally, the raw feature vector is converted to the log domain.

The exact definition of the feature extraction process is as follows: The FFT of the k-th sub-band of the audio signal is returned as vector of the form $\{x_{k,1}, x_{k,2}, \ldots, x_{k,N}\}$ and is *sorted into descending order of magnitude*, such that $x_{k,1} > x_{k,2} > \ldots > x_{k,N}$. The equations for calculating the spectral contrast feature from this sorted vector are as follows:

$$Peak_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i}\right) \quad (1)$$

$$Valley_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,N-i+1}\right) \quad (2)$$

and their difference is given by:

$$SC_k = Peak_k - Valley_k \quad (3)$$

where N is the total number of FFT bins in the k-th sub-band. α is set to a value between 0.02 and 0.2, but does not significantly affect performance. The raw Spectral contrast feature is returned as 12-dimensional vector of the form $\{SC_k, Valley_k\}$ where $k \in [1, 6]$. Although this feature is termed spectral contrast, suggesting that it is only the difference of the peaks and valleys, the amplitude of the spectral valleys are also returned to preserve more spectral information.

A signal that returns a high spectral contrast value will have high peaks and low valleys and is likely to represent a signal with a high degree of localised harmonic content. A signal that returns a low spectral contrast will have a lower ratio of peak to valley strength and will likely represent a signal with a lower degree of harmonic content and greater degree of noise components.

3.4. Marsyas-0.1 single vector Genre feature set

This dataset has also been classified by the *Genre* feature set included in Marsyas-0.1 [10], which estimates a single feature vector to represent a complete piece instead of a vector for each 23 ms of audio. This feature set includes beat, multi-pitch and timbral features in addition to MFCCs. The accurate comparison of algorithms in this type of research is difficult as there are currently no established test and query sets, however the inclusion of the *Genre* feature set allows comparison between "bag of frames" classifiers and classifiers which average spectral characteristics across a whole piece.

3.5. Reducing covariance in calculated features

The final step in the calculation of a feature set for classification is to reduce the covariance among the different dimensions of the feature vector. In the calculation of MFCCs this is performed by a Discrete Cosine Transform (DCT) [13]. However in [14] the calculation of spectral contrast feature makes use of the Karhunen-Loeve Transform (KLT), which is guaranteed to provide the optimal de-correlation of features. Both [14] and [15] suggest that the DCT is roughly equivalent to the KLT in terms of eliminating covariance in highly correlated signals. The de-correlated data from both transformations is output as a set of coefficients organised into descending order of variance, allowing us to easily select a subset of the coefficients for modelling, which include the majority of the variance in the data. This is known as the energy compaction property of the transformations.

3.6. Modelling temporal variation

Simple modelling of the temporal variation of features can be performed by calculating short time means and variances of each dimension of the calculated features at every frame, with a sliding window of 1 second. These means and variances are returned instead of the raw feature vector and encode a greater portion of the timbral information within the music. It is thought that this additional information will allow a classifier to successfully separate some styles of music which have similar spectral characteristics, but which vary them differently. This temporal smearing of the calculated features also spreads the meaningful data in some analysis frames across multiple frames, reducing the number of frames which do not encode any useful information for classification.

4. CLASSIFICATION SCHEME

In this evaluation musical audio signals were classified in to one of six genres, from which all of the test samples were drawn. The audio signals were converted into feature vectors, representing the content of the signal, which were then used to train and evaluate a number of different classifiers. The classifiers evaluated were single Gaussian models (with Mahalanobis distance measurements), 3 component Gaussian mixture models, Fisher's Criterion Linear Discriminant Analysis and new classifiers based on the un-supervised construction of a binary decision tree classifier, as described in [8], with either a linear discriminant analysis [9] or a pair of single Gaussians with Mahalanobis distance measurements used to split each node in the tree. We have only evaluated the performance of 3 component Gaussian mixture models because our initial results showed little improvement when the number of components was increased to 6, however the amount of time required to train the models increased significantly.

4.1. Classification and Regression Trees

In [8] maximal binary classification trees are built by forming a root node containing all the training data and then splitting that data into two child nodes by the thresholding of a single variable, a linear combination of variables or the value of a categorical variable. In this evaluation we have replaced the splitting process, which must form and evaluate a very large set of possible single variable splits, with either a linear discriminant analysis or a single Gaussian classifier with Mahalanobis distance measurements.

When using either linear discriminant analysis or a single Gaussian, to split a node in the tree, the set of possible splits of data is either the set of linear discrimination functions or the set of pairs of single Gaussians calculated from the set of possible combinations of classes. Therefore in this implementation, when a node in the classification tree is split, all the possible combinations of classes are formed and either the projections and discriminating points calculated or a single Gaussian is calculated for the two groups. Finally, the success of each potential split is evaluated and the combination of classes yielding the best split is chosen.

4.1.1. Selecting the best split

There are a number of different criterion available for evaluating the success of a split. In this evaluation we have used the Gini index of Diversity, which is given by:

$$i\left(t\right) = 2p\left(i|t\right)p\left(j|t\right) \tag{4}$$

where t is the current node, $p\left(j|t\right)$ and $p\left(i|t\right)$ are the prior probabilities of the positive and negative classes. The best split is the split that maximises the change in impurity. The change in impurity yielded by a split s of node t ($\Delta i\left(s,t\right)$) is given by:

$$\Delta i\left(s,t\right) = i\left(t\right) - P_L i\left(t_L\right) - P_R i\left(t_R\right) \tag{5}$$

where P_L and P_R are the proportion of examples in the child nodes t_L and t_R respectively. The Gini criterion will initially group together classes that are similar in some characteristic, but near the bottom of the tree, will prefer splits that isolate a single class from the rest of the data.

We have also examined the performance of the Twoing criterion [8] for evaluating the success of a split. Our results show that the performance of this criterion was nearly identical to that of the Gini criterion, which [8] suggests is because the performance of a classification tree is largely independent of the splitting criterion used to build it. In our initial experiments the performance of the Gini splitting criterion was often very slightly higher than that of the Two-ing criterion, hence the Gini criterion has been used in all subsequent evaluations.

4.1.2. *Building right sized trees and pruning*

In [8] it is shown that defining a rule to stop splitting nodes in the tree, when it is large enough, is less successful than building a maximal tree, which will over-fit the training set, and then pruning the tree back to a more sensible size. The maximal tree is pruned by selecting the weakest non-terminal node in the tree and removing its subtrees. The weakest link in the tree is selected by calculating a function G for each non-terminal node in the tree. G is formulated as follows:

$$G(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}, t \notin \tilde{T} \quad (6)$$

where $R(t)$ is the re-substitution estimate of node t as a leaf node, which is the misclassification cost of the training data if classified by majority vote at node t, $R(T_t)$ is the re-substitution estimate of the tree rooted at node t, $\tilde{T}_t$ is the set of all terminal or leaf nodes in the tree T_t and $|\tilde{T}_t|$ is the number of leaf nodes in the tree T_t. The node that produces the lowest value of G in the tree is identified as the weakest link, the whole tree is duplicated and the child nodes of the weakest node are removed. This process is continued until the root node is reached, yielding a finite, nested sequence of pruned trees, ranging from the maximal tree to the tree containing only the root node.

Once a finite, nested sequence of pruned trees has been produced, each tree is evaluated against an independent test sample, drawn from the same distribution as the training data. This allows us to identify trees that over-fit their training data, as they should return a higher miss-classification rate than the right-sized tree. Initially the tree with the lowest test sample estimate is selected. In order to reduce instability in the selection of the right sized tree, from a series of trees that may have very similar test sample estimates, the standard error (SE) of the test sample estimate is calculated and the simplest tree (smallest number of leaf nodes) within 1 standard error of the lowest scoring tree is selected as the output tree. The standard error is calculated as follows:

$$SE = \frac{R_{ts}(T)(1 - R_{ts}(T))}{N} \quad (7)$$

where $R_{ts}(T)$ is the independent test sample estimate of the misclassification cost of tree T and N is the number of examples in the test set.

5. TEST DATA AND CLASSIFICATION RESULTS

In this evaluation, we have used six classes of audio, each represented by 150 samples, which were a 30 second segment chosen at random from a song, also chosen at random from a database composed of audio identified by the authors as being from that genre of music. The first 10 seconds of each piece is ignored as this sometimes contains little data for classification. The genres selected were Rock, Classical, Heavy Metal, Drum and Bass, Reggae and Jungle music. Parameterisation of this data set yields approximately 1.2 million analysis frames for training and evaluation. Each experiment was performed 25 times and at each iteration, 50% of the data was chosen at random to be used for testing, whilst the other 50% of the data was used for training.

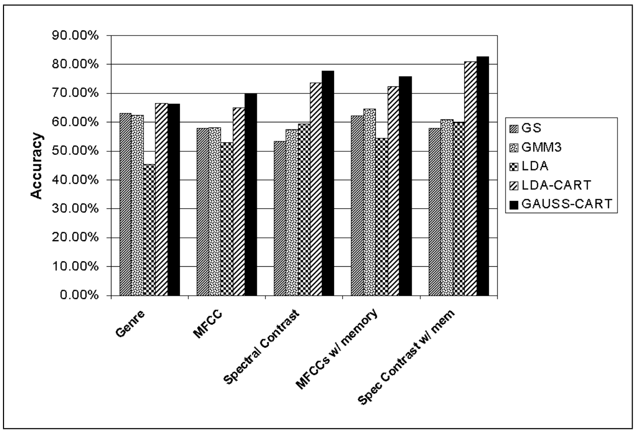

Figure 1. Bag of frames classification accuracies

The styles of music used in this evaluation have been deliberately chosen to produce a reasonably challenging dataset for evaluation. Jungle music is considered to be a sub-genre of Drum and Bass and is therefore quite similar to it and Heavy Metal is often considered to be a sub-genre of Rock music and so we should expect to see some confusion between these two genres. Heavy Metal can also be considered to be spectrally similar to Drum and Bass, as they have similar ratios of harmonic to non-harmonic content and percussive styles. Reggae can often be spectrally similar to Rock music, however the genres are melodically and rhythmically very different. It should also be noted that samples from Reggae music are often used in Jungle records, and that both the pace and style of the vocal parts in the two genres is almost identical, however the tempos of the drum tracks in Jungle music are 2 - 4 times as fast as those in Reggae.

In the figures, results labelled as GS correspond to the single Gaussian models, GMM to Gaussian mixture models, LDA to Fisher Criterion Linear Discriminant Analysis, LDA-CART to Classification trees with linear discriminant analysis and GAUSS-CART to Classification trees with single Gaussians and Mahalanobis distance measurements.

The "bag of frames" classification results in figure 1 show that there is little accuracy bonus to be gained through the use of Spectral Contrast feature instead of Mel Frequency based features. However, when used in conjunction with the decision tree classifier, the increase in classification accuracy over the Mel-frequency features is highly significant (8% for both the raw feature vectors and the temporally modelled feature vectors).

Temporal modelling of features increases the classification accuracy of MFCCs by 2 - 6% for flat classification schemes and 6 - 7% for the decision tree classifiers. The accuracy increase achieved for Spectral contrast features was 0 - 4% for flat classification schemes and 5 - 8% for the decision tree classifiers

In almost every case the decision tree classifier has achieved the greatest increases and has performed better than other models in accuracy, achieving increases of upto 12% and 21% for the raw MFCCs and temporally modelled MFCCs respectively, over Gaussian Mixture models. The increases achieved for raw Spectral Contrast feature and the temporally modelled version are 20% and 21% respectively.

The decision tree classifier based on single Gaussians has consistently performed better than the Linear Discriminant Analysis based classifier. However, it is interesting to note that in our initial experiments the individual frame classification accuracy is actually higher for the Linear Discriminant analysis based classifier in almost every case. Therefore, confusion must be better spread from the Gaussian based classifier in order to yield the greater "bag of frames" classification result.

When the results for "bag of frames" classification are compared to the single vector Genre feature extractor included in Marsyas-0.1 [10], it is clear that when using flat classifying schemes, accuracy with the Genre feature set is roughly equal to the accuracy achieved by Spectral contrast feature with temporal modelling. The decision tree classifiers yield a 4% improvement to the Genre feature set's accuracy, however Spectral Contrast feature with temporal modelling and a decision tree classifiers beats this by over 16% at 82.79% classification accuracy.

6. CONCLUSIONS

The separation of Reggae and Rock music was a particular problem for the feature extraction schemes evaluated here, perhaps because they not only share similar spectral characteristics but also similar ratios of harmonic to non-harmonic content, resulting in virtually no increase in accuracy for Spectral Contrast feature. The calculation of means and variances of the features helped to alleviate this confusion, perhaps by capturing some small amount of rhythmic variation in the one second temporal modelling window. Rock is a form of popular music with a heavily accented beat [1] whilst Reggae is a style of music with a strongly accented subsidiary beat [2], therefore, in order to completely separate Rock and Reggae music we would need to identify and separate the main and subsidiary (On and Off) beats, which would require a greater level of rhythmic modelling than is performed here, however this maybe approximated by the simple temporal modelling.

Similar trends are evident in classification of "Drum and Bass" music and "Jungle" music. Jungle music is closely related to Drum and Bass Music and is considered to be a sub-genre of Drum and Bass music as it has similar instrumentation and conforms to the same basic set of rhythmic rules, but imposes certain additional rhythmic restrictions. Temporal modelling of these genres achieves an increase in group classification accuracy but no increase in the separation of the two classes. This may be due to the absence of rhythmic modelling, as the two classes are often only differentiated by the length, complexity and repetition of the clearly defined rhythmic structures.

The large increases in accuracy achieved by the Classification and regression tree classifiers may be due to their ability to represent much more complex distributions within the data. Because the audio frames in this evaluation are quite short (23ms in length, which is close to the threshold of pitch/pulse perception) and the data is drawn from a complex, culturally based distributions, the distribution of each class in the feature space maybe very complex and interwoven with the other classes. The decision tree classifier allows the recursive division of the feature space into an unspecified number of tightly defined groups of sounds, which better represent the multi-modal distributions within the data. Effective classification is achieved by identifying groups of sounds which only occur in a certain class of music.

Gaussian models with a limited number of components are unable to model multi-modal distributions in the data. Increasing the separation of classes within the data by transformation can only be attempted once, and easily separable or outlier classes can cause other classes to be less well separated. By contrast, a decision tree classifier can perform different transformations at each level of the tree and is not limited by a fixed number of components.

6.1. McNemar's test

McNemar's test [16] is used to decide whether any apparent difference in error-rates between two algorithms, A_1 & A_2, tested on the same dataset is statistically significant. McNemar's test is performed by summarising the classification results of the two algorithms tested in the form of a two by two matrix containing the number of examples correctly classified by both algorithms(N_{00}), neither algorithm (N_{11}) and those only classified correctly by one of the algorithms (N_{10} & N_{01}). As there is no information about the relative performance of the two algorithms when they agree, these last two values are the only ones used in McNemar's test. Let H_0 be the hypothesis that the underlying error-rates are the same. Then under H_0 an error is as likely to be made by A_1 as A_2 and the distribution of N_{10} & N_{01} is the distribution obtained when tossing a fair coin and tails (N_{10}) is obtained. This is a binomial distribution and the P-values are easily obtained from tables.

McNemar's test has been applied to one iteration of each classification algorithm, with the same data and test sets. The results are summarised in figure 2. Results that have a P-value greater than 0.05 are not statistically significant and are shown in white, results with a P-value of 0.01 to 0.05 are shown in grey and statistically significant

[1] http://xgmidi.wtal.de/glossary.html

[2] http://simplythebest.net/music/glossary

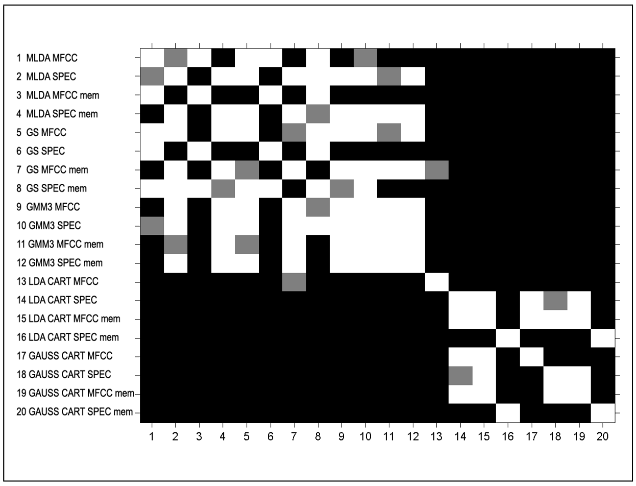

Figure 2. Statistical significance of classification results from McNemar's test

results, with a P-value of less than 0.01 are shown in black.

The algorithms in this figure have been grouped according to the classifier used. This shows a clear pattern in the results, the accuracy improvements made by the decision tree classifier are always statistically significant. Arranging the algorithms according to the feature set used or whether temporal modelling was used, produces no discernable pattern, other than a fragmented version of that produced by the classifiers. Clearly this indicates that the use of a decision tree classifier has had the most statistically significant effect on classification performance.

7. FURTHER WORK

Work in the future will concentrate on investigating methods of increasing the accuracy of these classifiers, including: calculating a confidence score for each classified frame and weighting the contribution to final classification by that score, selecting frames for classification, using variable frame rate or segmentation of the audio signal through onset detection and either including rhythmic analysis to the feature set or by adding categorical, rhythmic variable splits to the classification trees.

8. REFERENCES

[1] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals", *IEEE Transactions on Speech and Audio Processing*, 2002.

[2] George Tzanetakis, Georg Essl, and Perry Cook, "Automatic musical genre classification of audio signals", in *Proceedings of ISMIR 2001: The International Conference on Music Information Retrieval and Related Activities*.

[3] Pedro J Ponce de León and José M Iesta, "Feature-driven recognition of music styles", Tech. Rep., Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Ap. 99, E-03080 Alicante, Spain, 2003.

[4] Paul Scott, "Music classification using neural networks", Tech. Rep., Stanford University, Stanford, CA 94305, 2001.

[5] Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian, "Musical genre classification using support vector machines", Tech. Rep., Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613, 2003.

[6] Igor Karpov, "Hidden Markov classification for musical genres", Tech. Rep., Rice University, 2002.

[7] 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*.

[8] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone, *Classification and Regression Trees*, Wadsworth and Brooks/Cole Advanced books and Software, 1984.

[9] Andrew Webb, *Statistical Pattern Recognition*, John Wiley and Sons, Ltd, 2002.

[10] George Tzanetakis, "Marsyas: a software framework for computer audition", Web page, October 2003, http://marsyas.sourceforge.net/.

[11] S B Davis and P Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980.

[12] Alan P Schmidt and Trevor K M Stone, "Music classification and identification system", Tech. Rep., Department of Computer Science, University of Colorado, Boulder, 2002.

[13] N Ahmed, T Natarajan, and K Rao, "Discrete cosine transform", *IEEE Transactions on Computers*, 1974.

[14] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature", Tech. Rep., Department of Computer Science and Technology, Tsinghua University, China and Microsoft Research, Asia, 2002.

[15] Miodrag Potkonjak, Kyosun Kim, and Ramesh Karri, "Methodology for behavioral synthesis-based algorithm-level design space exploration: DCT case study", in *Design Automation Conference*, 1997, pp. 252–257.

[16] L Gillick and Stephen Cox, "Some statistical issues in the comparison of speech recognition algorithms", in *IEEE Conference on Accoustics, Speech and Signal Processing*, 1989, pp. 532–535.