Composer Identification of Digital Audio Modeling Content Specific Features Through Markov Models

Aric Bartle (abartle@stanford.edu)

December 14, 2012

1 Background

The field of composer recognition has been relatively well studied, and there are generally two ways to view this field. The first is to extract individual pitches along with their rhythm and dynamics from an exact source. Such a source can be a MIDI. This view has extensively been analysed through numerous models. Simpler models consist of training an SVM on the features of the composition that are the chords, the rhythms, the dynamics, etc.[LCY]. However, compositions are really much closer to actual language, and this implies that a NLP approach may be better. Much research takes this approach and uses such NLP models like Markov models successfully[JB].

The secondary view to composer classification is taken from the standpoint of the digital audio (waveform) of compositions. In a sense there is far less that has been explored from this standpoint. The Music Information Retrieval Evaluation eXchange (MIREX) is an annual evaluation campaign for composer identification and other tasks based on audio clips. Fairly impressive results are demonstrated each year, yet the classifications are done almost solely upon spectral analysis of the audio signal (spectral features); content specific features like actual harmonies are not considered in any depth. The paper [MU] analysed content specific features and obtained approximately a 5% accuracy increase over standard spectral analysis methods. This is not surprising since composers are mainly distinguished by musical content.

2 Overview

The system described here can be viewed as a linear SVM classifier with its features consisting of spectral and content specific ones. Several different types of content specific features are tried including the ones described in [MU] and a novel set of features generated through a Markov chain with the aim to see if these features produce better results than the ones in [MU].

The test and training sets for the SVM are drawn from a custom database consisting of 400 distinct audio samples of 30 seconds in length. The database is organized into 4 classes for the 4 different composers with each class being broken down into 4 albums of 25 samples. Here, the composers Bach, Beethoven, Chopin, and Liszt were chosen. All samples were distinct musical pieces and each album was a different performer. It should also be noted that these samples were restricted to the piano to avoid possibly biasing towards an instrument. In drawing the training and test samples, it is necessary that no sample from a particular album in the test set appear in the training and vice versa. This is to avoid the so called album effect where spectral features pick up on qualities relating to the recording and artist rather than the piece of music[JD].

3 Spectral Features

The spectral features are computed like in [GT]. However, only what are termed the timbral features are used in this implementation. Briefly they are Spectral Centroid, Rolloff, Flux and Mel-Frequency Cepstral Coefficients of an audio sample; in total, they form a 16 dimensional vector. This vector is computed throughout the audio sample and a running mean and running standard deviation are found as well. This creates for, when the two are combined, a 32 dimensional vector which can be extended to 64 by again computing a running mean and running standard deviation. Finally, these 64 dimensional vectors throughout the sample are averaged, giving a 64 dimensional spectral feature vector.

4 Content Specific Features

There are three different types of content feature vectors that are extracted. The first two involve initially estimating a sequence of harmonies (chords) throughout a sample, and then using those extracted harmonies to generate features. The harmonies consist of the 24 major and minor chords. Initially, an approach like in [MU] was used to extract these harmonies. However, it was found that there was quite a bit of noise in the predictions. Instead the software package [NMRB] was used to give fairly accurate predictions.

Lastly, these harmonies are transformed to the key of C major in order to be key invariant.

The first type of feature vector is computed like in [MU]. Transitions from one harmony to the next are used to form a 48 dimensional vector. The second type of feature is found through a Markov chain. A Markov chain is generated for each composer of the training set with the states being the harmonies. Then the log likelihood of each audio sample in the training set and test set are computed using these 4 Markov chains, yielding a 4 dimensional feature vector. The final type of feature vector is formed based upon the dynamics of a sample. The sample is first normalized with respect to the RMS of its corresponding album. This make the assumption that each album displays the full variation in dynamic range of that composer. After normalization, beat detection is performed in order to determine the likely locations where dynamics will change. In Western music, dynamics usually change on the beat. Finally, a max is taken around the beat and the resulting amplitude discretized, resulting in 1 of 5 levels of loudness. Like the second type of feature vector a Markov chain is employed to generate a 4 dimensional vector where this time the states are given by the levels of loudness.

5 Results

The database is split 25% (4 albums), 75% (12 albums) for the test set and training set, respectively. There are 1816 possible combinations for this splitting making it a bit impractical to compute all and form an average of the classification results. Instead, 64 random sets are computed and classified to form an average. This classification was run several times producing results (figure 1) all within 1.% of each other. A baseline of 76% accuracy was achieved through the spectral features alone (bar 1). When enhanced with the Markov specific features, classification increased to 81% (bar 2). However, if instead the 48 dimensional content specific feature vector was appended to just the spectral features, 83% accuracy was achieved (bar 3). Furthermore, another configuration was tested with one album in the testing set and the rest in the training set. There are 16 possible sets and the average of these classifications is summarized in figure 2.

It can be seen that in figure 2 the Markov features (bar 2) do better than the 48 dimensional vector (bar 3). This discrepancy can likely be explained by the nature of the training and test sets. In the first configuration, there can be only one album for a composer in the training set. Hence, the Markov model will be based off of just this album. It had been observed that the Markov model seemed to over fit by running tests on just the Markov features. It is reasonable to see that the Markov model produce worse results in this case and other similar ones. Another possible issue is that not all training/test set combinations were tried for the first classifier or enough trials performed. It was inherently prohibitive in terms of time to try all the 1816 combinations. Even performing the 64 trials took a considerable amount of time because of the relatively large regularization coefficient required.

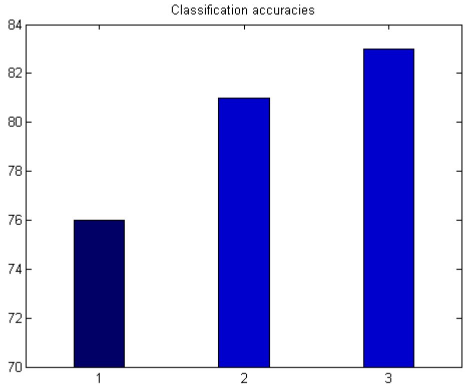

Figure 1: 4 to 12 split

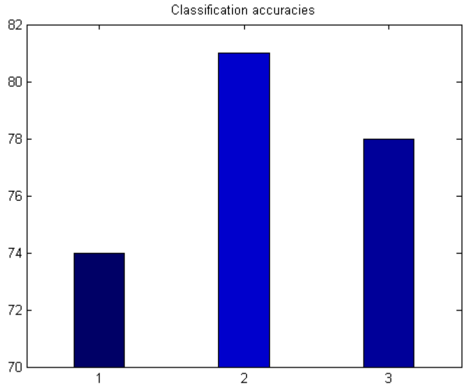

Figure 2: 1 to 15 split

6 Conclusion

The results are certainly promising. The second configuration shows the Markov features outperforming the one described in [MU] by 3%. The first configuration, although showing a decrease for the Markov features, cannot be taken that definitively because of the relative lack of testing, and even if in the end the Markov features do not perform as well, it is likely that the accuracy is not that less and above that of the spectral features alone.

It is certainly worthwhile to extend the Markov chain to more complex models. Furthermore, the content specific feature vectors can be extended as well to include information relating to the rhythm and such. Nonetheless, the underlying problem present is that for these more complex models the amount of data required grows quickly, and the time for constructing high quality audio clip databases grows quickly as well due mainly to the album effect. Fortunately, content specific features are indifferent to many of the problems that plague spectral features, particularly the album effect. It is because of this that future research should be capable of training and testing more complex models for content specific features.

References

[JB] Jan Buys. *Generative Models of Music for Style Imitation and Composer Recognition.*

[JD] J.S. Downie. *The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research.* Acoustic Science and Technology, 29, vol. 4, 2008.

[GT] George Tzanetakis. *MARSYAS SUBMISSIONS TO MIREX 2012.* http://www.music- ir.org/mirex/abstracts/2012/GT1.pdf.

[LCY] Justin Lebar, Gary Chang, David Yu. *Classifying Musical Scores by Composer:A machine learning approach*

[MU] Sean Meador, Karl Uhlig. *Content-Based Features in the Composer Identification Problem CS 229 Final Project.*

[NMRB] Yizhao Ni, Matt Mcvicar, Raul Santos-Rodriguez, Tijl De Bie. *Harmony Progression Analyser for harmonic analysis of music.* https://patterns.enm.bris.ac.uk/hpa-software-package.