Elliot Sinyor
119931368

Musical Genre Similarity

Written Summary of Oral Presentation for MUMT 611

Introduction

With the number of online music files in the millions and steadily increasing, there is still no standardized way to automatically classify pieces by genre, nor is there even a standard set of genre descriptors. In fact, there are some inherent problems in that there is significant overlap between genres, and the boundaries are often fuzzy at best.

In their 2003 paper *Representing Musical Genre: A state of the art*, Aucouturier and Pachet outline the three main approaches to genre classification, namely manual, prescriptive and emergent. Manual classification refers to using human experts to classify pieces of music. Of the automatic approaches, prescriptive refers to the use of low-level audio features of the sound, whereas emergent refers to the use of existing text meta-data available in online databases. This paper will briefly present some related work, and then describe two automatic approaches.

Related Work

Work on automatic genre classification stemmed from speech recognition research. Low-level audio features such as zero-crossing rate and mel-frequency cepstral coefficients (MFCC) have been used in systems that distinguish between music, speech and environmental noise. Most of the approaches were motivated by the need to classify hours of audio coming from broadcast TV. In (Kimber et al. 1996), the authors used MFCCs and then an HMM to classify signals into music, speech, laughter and non-speech. Zhang and Kuo present a system that classifies audio from movies and TV into two main classes, music and non-music. Within *music* it distinguishes between harmonic environmental sound, pure music, song, speech with music, and environmental sound with music. For *non-music* it can distinguish between pure speech and non-harmonic environmental sound. In (Soltau et al. 1998), the authors describe a scheme that uses a modified neural network to classify audio signals as either Rock, Pop, Classical or Techno based on cepstral coefficients. They compare their enhanced neural network with an HMM-based classification scheme for 360 30-second samples.

Manual Approaches

Manual approaches to musical genre classification involve musicologists listening to samples and then classifying them. In Dannenberg et al. 2001, one of the authors describes Microsoft's attempt at classifying several hundred thousand songs for their MSN Music Search Engine. They hired full-time musicologists and it took approximately 30 human-years to classify the music.

Aucouturier and Pachet also briefly describe their efforts to develop a genre taxonomy for the CUIDADO project. They assert that while manual classification is unfeasible for the millions of pieces of music currently online, it remains a useful way to develop a taxonomy or as a means of evaluating automatic algorithms.

Automatic (Prescriptive) Approach:
Tzanetakis and Cook, *Musical Genre Classification of Audio Signals*, 2002.

In their paper, Tzanetakis and Cook describe a classification scheme based on timbral features, rhythmic content and pitch content. After assembling a feature vector encompassing the above-mentioned feature sets, they compared the performance of various statistical pattern-recognition methods.

The timbral features used are spectral centroid, spectral flux, zero-crossing rate, and five MFCCs. These features are measured over a short time frame (a 23 ms *analysis window*) and then their means and variances over 43 *analysis windows* are combined to form the timbral-texture feature vector. An additional feature called the *low energy feature* is also included as it can indicate that several *anaylsis windows* had below-average energy, which can possibly distinguish between genres.

To quantify the rhythmic content of each piece, the authors devised a measure called a *beat histogram* (BH) and included features derived from the BH in a rhythmic content feature vector. The BH is made by first separating the signal into a number of frequency bands using the discrete wavelet transform. Then, after full-wave rectifying, lowpass-filtering and downsampling each band, an autocorrelation function is applied to all bands, and a peak-picking algorithm is used to construct a histogram. They liken the procedure to "pitch detection with larger periods". One the BH is created, they derive several features to be added to the overall feature vector. Theses features are: the amplitudes of the first and second histogram peaks, the ratio of the amplitudes of the two highest peaks, the periods of the first and second peaks in BPM, and the overall sum of the histogram.

Similar to the beat histogram, the final feature set is derived from a *pitch histogram* (PH). Here the signal is divided into two frequency bands, and amplitude envelopes are extracted for each band. Then, using an enhanced autocorrelation function, the three dominant peaks are accumulated into a PH for the whole file. In fact, two PHs are found, a folded one (FPH) representing the pitch classes, and an unfolded one (UPH) representing the pitch range. The features derived from the PHs include the maximum amplitude of the FPH, the period of the maximum *unfolded* peak, the period of the maximum *folded* peak, the ratio of the two highest *folded* peaks, and the overall sum of the histograms.

Various classification schemes were compared, including a simple Gaussian classifier, Gaussian mixture models using the K-means algorithm with various values of K, and a K-nearest neighbour classifier again with various values of K. The dataset, consisting of 19 hours of audio data (100 samples * 23 genres * 30 sec) was randomly partitioned so that 90% is used for training and 10% is used for testing. The main genres used are Classical, Country, Disco, Hip Hop, Jazz, Rock, Blues, Reggae, Pop and Metal. Classical is subdivided into Choir, Orchestra, Piano, and String Quartet. Jazz is divided into Big Band, Cool, Fusion, Piano, Quartet and Swing. The best case results were for a GMM with K = 3 distributions. In this case, accuracy was found to be 61% for the main genre categories, 88% for classical categories and 68% for Jazz categories.

Automatic (Emergent) Approach:
Pachet, Westermann, and Laigre, *Musical Data Mining for Electronic Music Distribution*, 2001

In this paper, the authors describe an emergent approach to classification, namely co-occurrence analysis. The paper begins by mentioning another emergent approach, collaborative filtering, which refers to using subjective ratings entered by users to find other users with similar tastes and finding artists tracks that each user might enjoy. Co-occurrence analysis, on the other hand, refers to the use of online lists to find instances of co-occurrence between two titles. The reasoning is that "if two items appear in the same context, that is evidence that there is some kind of similarity between them." The lists used in the paper were radio playlists from Radio France and compilation CD tracklistings from CDDB.

The heart of the approach lies in creating a matrix such that the value at (*i, j*) corresponds to the number of times that title *i* co-occurs with title *j* for a pre-determined bank of text lists. One problem that immediately arises with this approach is that often two titles may not co-occur directly, but might have a common neighbor. The authors describe a correlation distance function that takes such indirect co-occurrences into account.

Using both co-occurrence and correlation distance functions, the distances are clustered using Ascendant Hierarchical clustering. For their set of 100 artists, they find a 76% accuracy rate using CDDB and 70% for Radio France for co-occurrence clustering for level 1 clusters, meaning groups of two artists. Correlation clustering, on the other hand, yields 53% for Radio France and 59% for CDDB. For higher level clusters, correlation clustering yields accuracy rates of 47% for Radio France and 74% for CDDB. Co-occurrence clustering yields 28% for Radio France and 54% for CDDB. From these results, the authors affirm that correlation clustering indicates that items in a bigger cluster are likely similar in genre, whereas smaller co-occurrence clusters indicate similarity between two titles.

References

Aucouturier J., and F. Pachet. Representing musical genre: A state of the art. 2003. *Journal of New Music Research*.

Tzanetakis G., and P. Cook. Musical genre classification of audio signals. 2002. *IEEE Transactions on speech and audio processing*.

Pachet F., G. Westermann, and D. Laigre. Musical data mining for electronic music distribution. 2001. *Proceedings of the International Conference on Web Delivering of Music*.

Soltau H., T. Schultz, M. Westphal, and A. Waibel. Recognition of music types. 1998. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*

Li T., M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. 2003. *Proceedings of the ACM conference on research and development on information retrieval.*