Indian Institute of Technology, Guwahati

BTP REPORT: PHASE 1

Automatic Genre Classification of Audio Signals

Abhinav Singh & Ashna Dhanda
Supervisor: Professor S.R.M. Prasanna

Bachelor of Technology Program,
Electronics and Communication Department,
Indian Institute of Technology,
Guwahati, Assam, India

Abstract:

The creation of huge databases coming from both restoration of existing analog archives and new content is demanding more and more reliable and fast tools for content analysis and description, to be used for searches, content queries and interactive access. In that context, musical genres are crucial descriptors since they have been widely used for years to organize music catalogues, libraries and music stores. Despite their use, musical genres remain a poorly defined concept, which make of the automatic classification problem a non-trivial task.

In this report, we review the state-of-the-art in automatic genre classification and content description and present new concepts and algorithms under development for real world databases. As description of audio is a broad field that incorporates many techniques, an overview of the main directions in current research is given. However, a detailed study of automatic audio classification is conducted and a music genre classifier is designed. To evaluate the classifier, a general database is created comprising of about 2000 songs from 9 different musical genres.

The classification algorithm used for the development of music genre classifier is Vector Quantization, which is commonly used for the task. Based on feature's effectiveness, a robust musical genre classifier is designed and a classification accuracy of 86.77% is achieved over audio files from 9 different musical genres

1 Introduction

1.1 Background

Our daily life is highly dependent on information, for example in formats as text and multimedia. We need information for common routines as watching/reading the news, listening to radio, watching a video etc. However, we easily run into problems when a certain type of information is needed. The immense flow of information makes it hard to find what you are looking for.

The rapid increase of information imposes new demands of content management as the media archives and consumer products begin to be very complex and hard to handle. Currently people perform searches in databases with different meta-tags, which only describe whole chunk of information with a constellation of tags. The meta-tags are constructed by different individuals and one realizes that the interpretation of meta-tags can differ from individual to individual. An automatic system that describes audio would systematize labeling and would also allow searches on the actual data, not just on labels of it. Better content management is the goal of the automatic audio description systems. Some commercial content management applications are already out but many possible applications are still undeveloped. *For example*, a person is listening to the radio and wants to listen to jazz. Unfortunately, all the radio stations play pop music mixed with advertisements. The listener gives up searching for jazz and gets stuck with the pop music. This problem can be solved with an automatic audio description system. The scenario may then change to following. The person that wants to listen to jazz only finds pop music on all the tuned radio stations. The listener then presses a 'search for jazz' button on the receiver and after a couple of seconds the receiver change radio station and jazz flows out of speaker. This examples show how content management may be an efficient tool that simplifies daily routines involving information management.

Musical genres are categories that have arisen through a complex interplay of cultures, artists and market forces to characterize similarities between musicians or compositions and organize music collections. Yet, the boundaries between genres still remain fuzzy as well as their definition making the problem of automatic classification a non-trivial task. The Music Genre Classification problem asks for taxonomy of genres i.e. a hierarchical set of categories to be mapped onto a music collection. Pachet and Cazaly [2] studied a number of musical genre taxonomies used in industry and on the Internet and showed that it is not straightforward to build up such a hierarchy of genres. As a good classification relies on a carefully thought taxonomy, we start here from a discussion on a number of critical issues.

Music Genre can be classified by two ways: *Non - Musical Criteria* and *Musical criteria*. Music may also be categorized by non-musical criteria such as geographical origin though a single geographical category will normally include a wide variety of sub-genres. It can also be said that a music genre (or subgenre) is defined by the techniques, the styles, the context and the themes (content, spirit). We further discuss some of the important aspects associated with some musical genres around the world.

- *Classical Music (European Classical Music)*
  In common usage classical music often refers to orchestral music in general, regardless of when it was composed or for what purpose. A full size orchestra (about 104 players) may sometimes be called a "symphony orchestra". The typical orchestra consists of four proportionate groups of similar musical instruments, generally appearing in the musical score in the following order:
  Woodwinds: 2 flutes, 2 clarinets, bass clarinet, etc
  Brass: 5 trumpets, 2 to 6 horns etc.
  Percussion: Snare drum, bass drum, timpani etc.
  Strings: violins, harps, cellos, double basses, pianos etc.
  Flutes, clarinets, horns, trumpets, timpani, violins, cellos are among the core symphonic instruments.

- *Jazz Music*
  Jazz is a musical form that grew out of cross-fertilization of folk, blues and other music. Jazz is primarily an instrumental form of music. The instrument most closely associated with jazz may be the saxophone, followed by the trumpets. The piano, guitar and drums are also primary jazz instruments. The single most distinguishing characteristics of jazz are improvisation. Jazz also tend to utilize complex chord structures and an advanced sense of harmony.

- *Rock Music*
  Rock in its broadest sense can refer to almost all popular music recorded since 1950. Its main feature includes an emphasis on rhythm and the use of the electric guitars. Rock is a form of popular music from the late 20th century which typically features a vocal melody (often with vocal harmony) that is supported by accompaniment of electric guitars, a bass guitar, and drums, often with a strong back beat. Keyboard instruments such as organ, piano, or synthesizers are often used in many types of rock music. While brass instruments, such as saxophone were common in some styles in earlier development of rock, they are less common in the newer subgenres of rock music since the 1990s. The genre of rock music is broad, and its boundaries loosely-defined, with related genres such as soul and funk sometimes being included in the definition of the term.

- *Electronic Music*
  Electronic music is a term for music created using electronic devices. As defined by the IEEE standards body, electronic devices are low-power systems and use components such as transistors and integrated circuits. Working from this definition, distinction can be made between instruments that produce sound through electromechanical means as opposed to instruments that produce sound using electronic components. Examples of an electromechanical instrument are the teleharmonium, Hammond B3, and the electric guitar, whereas examples of an electronic instrument are a Theremin, synthesizer, and a computer.

Pachet and Cazaly [2] showed that a general agreement on genre taxonomies does not exist. Taking the example of well known websites like AllMusic (531 Genres), Amazon (719 Genres), and Mp3 (430 Genres), they only found 70 terms common to the 3 taxonomies. Furthermore, genre taxonomies may be dependant on cultural references. For example, a song by the French singer Charles Aznavour would be considered as "Variety" in France but would be filed as "World Music" in the UK. Due to difficulty of defining a universal taxonomy, more reasonable goals must be considered. In fact, Pachet and Cazaly eventually gave up their initial goal to define a general taxonomy of musical genres [2] and Pachet and al. decided to use simple two-level genre taxonomy of 20 genres and 250 subgenres in the context of the Cuidado music browser [3].

Musical genres are the main top-level descriptors used by music dealers and librarians to organize their music collections. Though they may represent a simplification of one artist's musical discourse, they are of a great interest as summaries of some shared characteristics in music pieces. With Electronic Music Distribution (EMD), music catalogues tend to become huge (the biggest online services propose around 1 million tracks); in that context, associating a genre to a musical piece is crucial to help users finding what they are looking for. In fact, the amount of digital music urges for efficient ways to browse, organize and dynamically update collections: it definitely requires new means for automatic annotation. In the case of music genre annotation, Weare [1] reports that the manually labeling of hundred thousand songs for Microsoft's MSN Music Search Engine needed about 30 musicologists for one year. At the same time, even if terms such as *jazz*, *rock* or *pop* are widely used, they often remain loosely defined so that the problem of automatic genre classification becomes a non-trivial task

1.2 Related Work

1.2.1 Feature Extraction

In the digital media world, generic audio information is mostly represented by bits allowing a direct reconstruction of an analogue waveform. In real world applications a precise symbolic representation of a (new) song is rarely available and one has to deal directly with most straightforward form, i.e. audio samples. Audio samples, though sampling the exact sound waveform, can not be used directly by automatic analysis systems because of the low level and low "density" of the information they contain; put in another way, the amount of data is huge and the information contained in audio samples taken independently is too small to deal with humans at the perceptual layer.

The first step of analysis systems is thus to extract some *features* from the audio data to manipulate more meaningful information and to reduce the further processing. Extracting features is the first step of most pattern recognition systems. Indeed, once significant features are extracted, any classification scheme may be used. In the case of audio signals, features may be related to the main elements of music including *melody, harmony, rhythm, timbre and spatial location.*

1.2.1.1 Timbral Features

Timbre is currently defined in literature as the perceptual feature that makes two sounds with the same pitch and loudness sound different. Features characterizing timbre analyze the spectral distribution of the signal though some of them are computed in the time domain. These features are global in the sense that they integrate the information of all sources and instruments at the same time. Most of these descriptors are computed at regular time intervals, over short windows of typical length between 10 and 30 ms.

- *Zero-Crossing Rate [4], [6]:* it is defined as the number of zero-crossings of the signal in the time domain; it is a measure of noisiness of the signal and it is correlated with pitch. Algorithm is simple and has low computational complexity. Scheirer use the ZCR to classify audio between speech/music, Tzanetakis [17] use it to classify audio into different genres of music and Gouyon use it to classify percussive sounds. Some part of music has variations in ZCR. For instance, a drum intro is a pop song can have high variation in ZCR values.
- *RMS [6]:* It is a measurement of the energy of a signal. The RMS value is however defined to be the square root of the average of a squared signal.
- *Loudness [6]:* simple models of loudness consist typically of an exponentiation of the energy of the frame.
- *Low Energy Feature [4], [6], [15]:* it is the percentage of frames within a larger window that have RMS energy lower than the mean RMS energy across the window.
- *Linear Prediction Coefficients [7], [8*]: linear prediction has been studied in the context of speech recognition to model sound production: the observed sound is supposed to be the result of the linear filtering of a simple signal. The estimated coefficients of the filter can be used as timbre descriptors since they encode the effect of the resonating body of the instrument (or of the vocal track in the case of speech production).
- *FFT coefficients [8]:* the feature vector is simply the vector of the FFT coefficients.
- *Mel-Frequency Cepstrum Coefficients (MFCCs) [4], [5], [6], [7], [9], [14], [15]:* they are perceptually motivated features obtained by taking the log-amplitude of the magnitude spectrum, warping the spectrum onto a perceptual frequency scale (the Mel frequency scale) and by applying a discrete cosine transform on the Mel coefficients to decorrelate the resulting feature vector. Thirteen coefficients are typically used for speech recognition.
- *Spectral Centroid [4], [6]:* it describes the center of gravity of the power spectrum. It is a cheap description of the shape of the power spectrum. It indicates whether an audio spectrum is dominated by low or high frequencies and additionally it is correlated with a major perceptual dimension of timbre; i.e. sharpness.
- *Spectral Roll-Off [4], [6]:* it is a measure of spectral shape. It is defined as the frequency below which most of the power spectrum is concentrated (typically 85 % of the power spectrum).
- *Spectral Flux [4], [6*]: it is a measure of the amount of local spectral change. It is evaluated as the difference between the normalized magnitudes of successive spectral distributions.
- *Spectrum Spread [6]:* it describes the second moment of the power spectrum. It is a cheap descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or spread out over the spectrum. It allows differentiating between tone-like and noise-like sounds.
- *Spectrum Flatness [6]:* it expresses the deviation of the signal's power spectrum over frequency from a flat shape (corresponding to a noise-like or an impulse-like signal). A high deviation from a flat shape may indicate the presence of tonal components.
- *Harmonic Ratio [6]:* it is loosely defined as the proportion of harmonic components within the power spectrum. It is derived from the correlation between the signal and a lagged representation of the signal, lagged by the fundamental period of the signal.
- *Wavelet [10]:* the wavelet decomposition scheme allows a subdivision of the signal into octave sub-bands while providing good time and frequency resolution. The Daubechies Wavelet Coefficients Histograms (DWCH's) proposed in [10] are histograms of wavelet coefficients over a large window from which features are extracted by evaluating moments.

Most of the proposed algorithms for musical genre classification use indeed one small segment of audio per title: typically a 30-second long segment starting 30 seconds after the beginning of the piece to avoid introductions that may not be representative of the whole piece. Extraction of high-level descriptors from unrestricted polyphonic audio signals is not yet state of the art. Thus most approaches focus on timbre modeling based on combinations of low-level descriptors. Timbre may contain sufficient information to roughly characterize musical genres as research demonstrated that humans with little to moderate musical training were able to perform a correct classification of music (among 10 genres) in 53 % of the cases after listening to only 250 milliseconds and in 72% of cases based on only 3 seconds of audio [12]. This suggests that no high-level understanding of music is needed to characterize genres as 250 milliseconds and in a lesser manner 3 seconds are too little time to recognize a musical structure.

1.2.2 Pattern Classifiers

Extracting features is the first step of most pattern recognition systems. Indeed, once significant features are extracted, any classification scheme may be used. In the case of audio signals, a number of classifiers have been used in the literature. We further describe a few of them in this report:

1.2.2.1 The Unsupervised Algorithms

In the last few years, the machine learning approach has garnered increasing interest. In the unsupervised approach, an audio title is represented by a set of features as seen in section 1.2.1 and a similarity measure is used to compare titles among each others. Unsupervised clustering algorithms take then advantage of the similarity measure to organize the music collection with clusters of similar titles.

Similarity Measures

The simplest choice to measure distance between two feature vectors is to use a Euclidean distance or a cosine distance for example. However these distances will only make sense if the feature vectors are time-invariant. Otherwise two perceptually similar titles may be distant according to the measure if the similar features are time-shifted. A possible solution to build a time-invariant representation of a time series of feature vectors is to firstly build statistical models of the distribution of the features and then use the distance to compare these models directly. Typical models include K-means, Gaussian and Gaussian mixtures (GMMs).

Clustering Algorithms

K-means is probably the simplest and most popular clustering algorithm. It allows partitioning a set of vectors into K disjoint subsets. One of its weaknesses is that it requires the number of clusters (K) to be known in advance. Shao et al. [7] cluster their music collection with the Agglomerative Hierarchical Clustering, a clustering algorithm that starts with *N* singleton clusters (where *N* is the number of titles of the database) and that forms a sequence of clusters by successive merging. The Self-Organizing Map (SOM) and the Growing Hierarchical Self-Organizing Map (GHSOM) are used to cluster data and organize them on a 2-dimensional space in such a way that similar feature vectors are grouped close together. SOMs are unsupervised artificial neural networks that map high dimensional input data onto lower-dimensional output spaces while preserving the topological relationships between the input data items as faithfully as possible.

In some terms, the major drawback of unsupervised techniques can be that the obtained clusters are not labeled. In any case, the obtained clusters do not always reflect genre hierarchies, rather similarities dependent on the type of features (rhythmical similarities, melodic similarities, etc.).

1.2.2.2 The Supervised Algorithms

The supervised approach to music genre classification has been studied more extensively. The methods of this group suppose that taxonomy of genres is given and they try to map a database of songs into it by machine learning algorithms. As a first step, the system is trained with some manually labeled data, and then it is used to classify unlabelled data.

We describe here a number of commonly used supervised machine learning algorithms. We do not pretend to make an exhaustive list of such algorithms but to focus on those that have been used in the context of music genre classification.

Supervised Classifiers

- *K-Nearest Neighbor (KNN):* it is a non-parametric classifier based on the idea that a small number of neighbours influence the decision on a point. More precisely, for a given feature vector in the target set, the K closest vectors in the training set are selected (according to some distance measures) and the target feature vector is assigned the label of the most represented class in the K neighbours. KNNs are evaluated in the context of genre classification in [13], [10], [4].
- *Gaussian Mixture Models (GMM):* GMMs model the distribution of feature vectors. For each class, the existence is assumed of a probability density function expressible as a mixture of a number of multi-dimensional Gaussian distributions. The iterative Expectation Maximization (EM) algorithm is used to estimate the parameters for each Gaussian component and the mixture weights. GMMs have been widely used in the music information retrieval community, notably to build timbre models as seen in above section. They have been used to model directly musical genres in [10], [4], [5]. In [6], a tree-like structure of GMMs is used to model the underlying genre taxonomy: a divide-and-conquer strategy is used to first classify items on a coarse level and then on successively finer levels. The classification decision is thus decomposed into a number of local routing or refinement decisions in the taxonomy. In addition, feature selection at every refinement level allows optimizing classification results
- *Support Vector Machines (SVM*): SVMs are based on two properties: margin maximization (which allows for a good generalization of the classifier) and nonlinear transformation of the feature space with kernels (as a data set is more easily separable in a high dimensional feature space). SVMs have been notably used in the context of genre classification by [10], [14], [15].
- *Hidden Markov Model (HMM):* Hidden Markov Models can also be used for classification purposes. They have been extensively used in speech recognition because of their capacity to handle time series data. HMMs may be seen as a double embedded stochastic process: one

process is not directly observable (hidden) and can only be observed through another stochastic process (observable) that produces the time set of observations. Though they may be well suited to modeling music, to our knowledge, HMMs have only been used in [9] and [14] for genre classification of audio content (they have been used in [7] as well but in the case of unsupervised organization of a music collection).

- *Mixture of Experts (ME):* A Mixture of Experts solves a classification problem by using a number of classifiers to decompose it into a series of sub-problems. Not only does it reduce the complexity of each single task but it also improves the global accuracy by combining the results of the different classifiers (experts). Of course, the number of needed classifiers is increased but having each of them a simpler problem to handle; the overall required computational power is reduced. Using a mixture of classifiers, each subtask may focus either on a subset of the attributes (feature selection), on different sample data (resampling, i.e. sub-sampling, bagging, boosting…), or on a different data labelling (decomposition of polychotomies into dichotomies). A range of solutions has been proposed in literature for the combination of different models into a global system. A possible solution is to use a majority vote of the different experts. This solution is used in [13] where 3 different MLPs characterize three different segments of a single song. Another possibility is to average the output probability estimates of each expert for each class. A more elaborated strategy is to consider the combination of the outputs of each expert as another learning problem. When each expert is working on a different feature set, the combination of results can be itself a function of the features so that inputs control the weights associated to each expert: such a method implements a sort of dynamic feature selection strategy. This latter classification scheme is used in [15] with SVM expert and three different feature sets.

2 Automatic Audio Genre Classification System

In this section we present the algorithms used to develop a robust Automatic Audio Genre Classifier for associating automatically a music genre to an audio excerpt. Our algorithm parameterizes audio content by extracting 2 set of features describing 2 different dimensions of music: timbre and energy. Once features extracted, Vector Quantization is used for classification into musical genres. The underlying idea is to use separate models to approximate different parts of a problem and to combine the outputs from the models finally.

This section is organized as follows: In 2.1 we discuss the feature extraction techniques used for extracting features from the audio excerpt. Then in section 2.2 we discuss the classifier used for classification of the feature sets.

2.1 Feature Extraction

2.1.1 Segmentation into analysis frames

The audio excerpts used are sampled at 44100 Hz, 16-bit resolution and converted to mono signals. The first 30 seconds of the signals are discarded to avoid introduction that may not be representative of the rest of the excerpt. Only the next 30 seconds of the signal are kept for further analysis to limit further processing. The resulting signals are then analyzed through sliding windows of 23 ms overlapped by 50%. In the case of genre classification, it is probable that these precision requirements could be relaxed. West and Cox [16] use audio signals sampled at 22050 Hz and no overlap between the frames without significant loss of the classification accuracy. Further experiments have to be run in our case to check if the system is robust to signals with reduced quality.

2.1.2 Texture windows

Frames of 23 ms are used for short time Fourier transforms analysis since they all representing the evolution of spectrum with a good precision. Yet this time scale too many variations occur. Some integration process must be held to build more robust features. Not only does it reduce further computations but it is also more perceptually relevant. Consequently, *texture windows* are used to combine low-level features of adjacent analysis frames.

The impact of the size of the window over classification accuracy has been studied in [17]. The conclusion is that texture windows of 1 second are a good compromise since no significant gain in classification accuracy is obtained by taking larger windows while the accuracy decreases as the window is shortened.

Scaringella [18] made experiments with the texture windows centered on the time positions of musical beats. The sizes of the corresponding windows were selected in accordance with the local beat rate of the excerpt. Though this may allow a perceptually more relevant modeling of musical signals, no significant improvement of classification accuracy has been obtained with this technique, probably because of the weakness of the state-of-the-art beat tracker. Consequently, the algorithm presented here use simple 1-second texture windows.

2.1.3 Timbre features

Mel-Frequency Cepstral Coefficients (MFCC's) are computed from the analysis frames. MFCC's are widely used descriptor for timbre modeling coming from the speech recognition literature [19]. Each analysis frame is parameterized with 20 MFCC's. The number of MFCC's used has been chosen to limit the further computations rather than by a careful analysis of its impact on the classification accuracy, though the number of MFCC's is a subject of debate in the literature [20]. Mean, Standard deviation over the texture window are evaluated for each MFCC resulting in 3 different feature vector sets each of 20 dimensions.

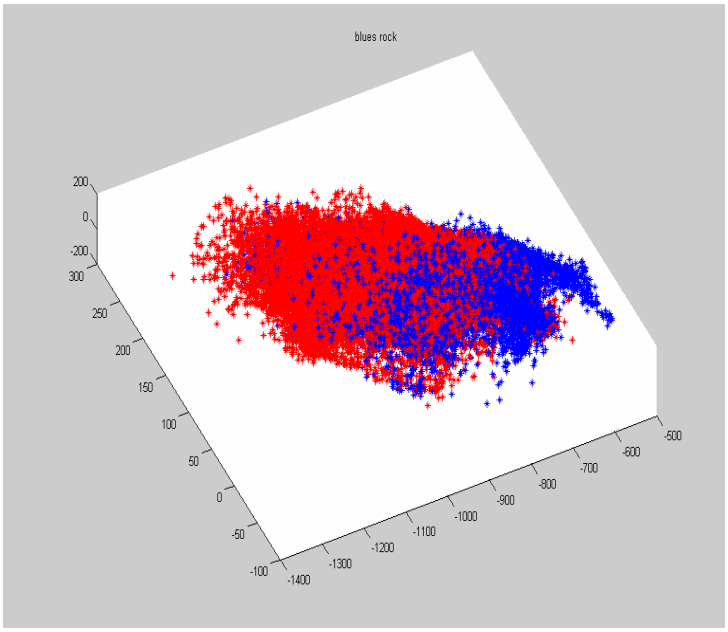

Figure1: Blues v/s Rock

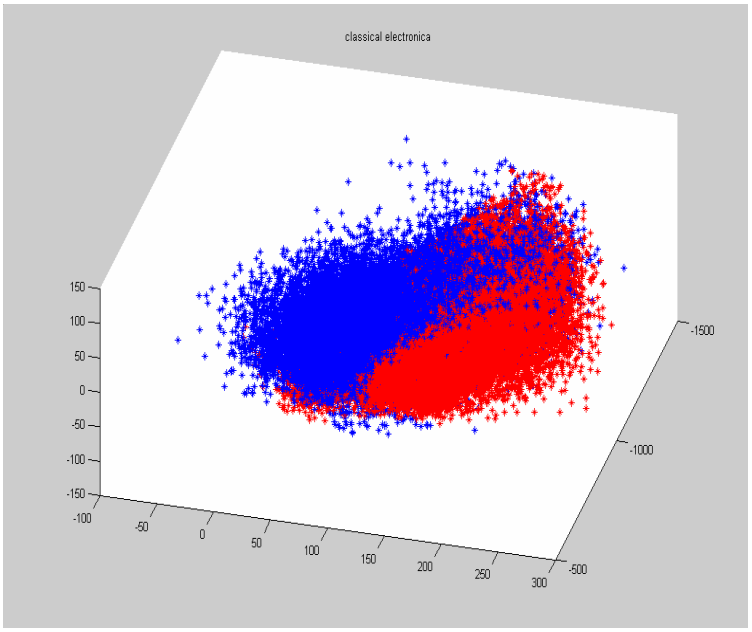

Figure3: Classical v/s Electronic

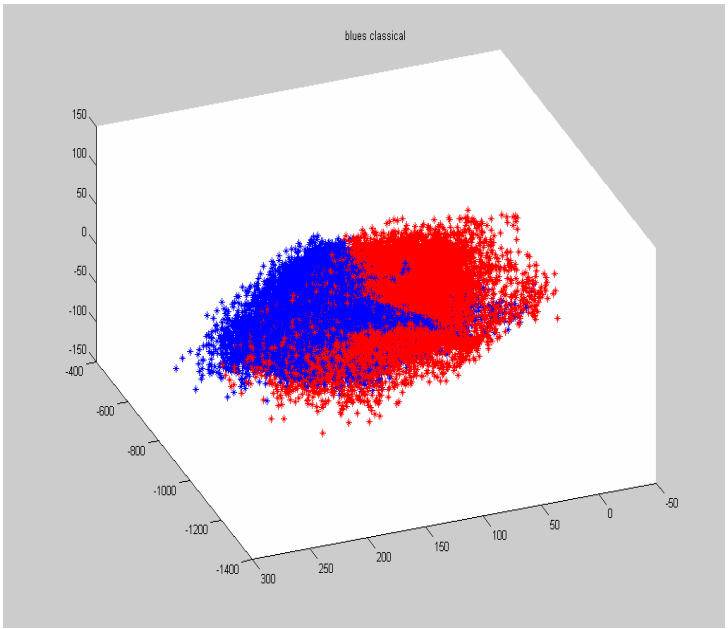

Figure2: Blues v/s Classical

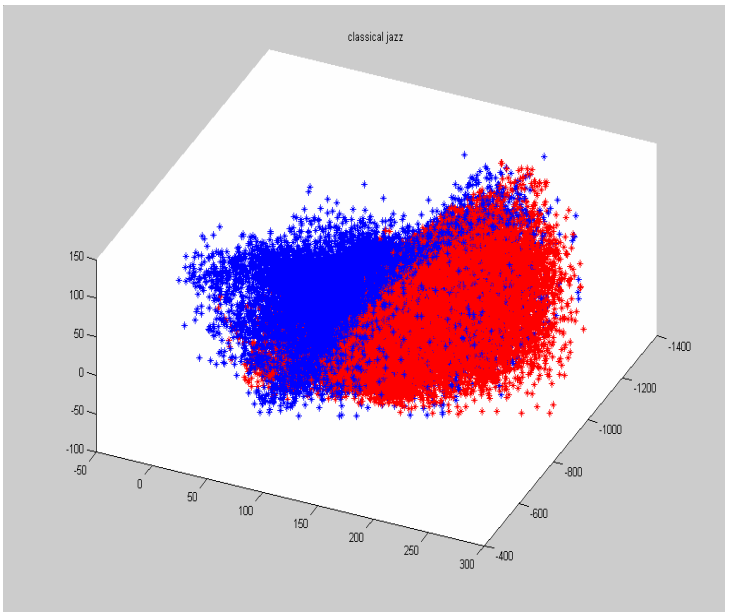

Figure4: Classical v/s Jazz

2.1.4 Energy Features

As we musician play music in a particular octave for a given song, we tried out extracting energy in 6 octaves that span from 62.5 Hz to 22050 Hz. Log-compressed energies in 6 frequency bands are extracted from each analysis frame. Each band covers roughly one octave. Mean. Standard deviations of each coefficient are evaluated over the texture window.

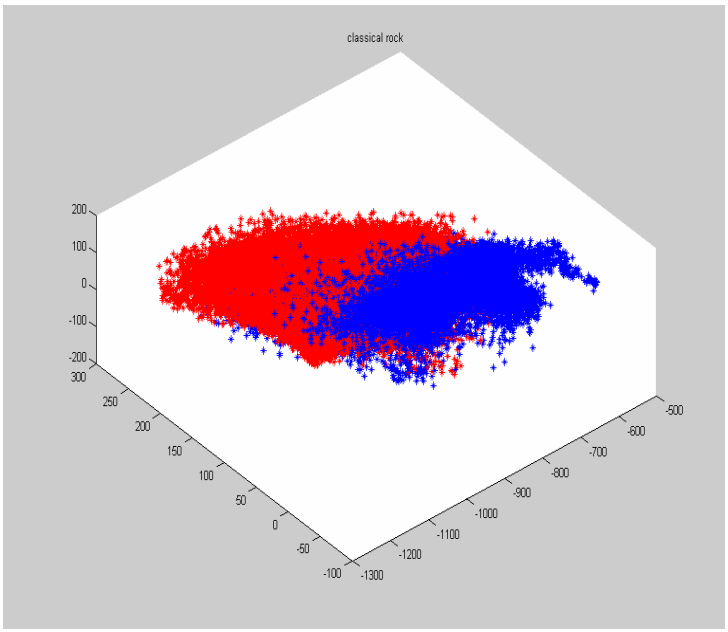

Figure5: Classical v/s Rock

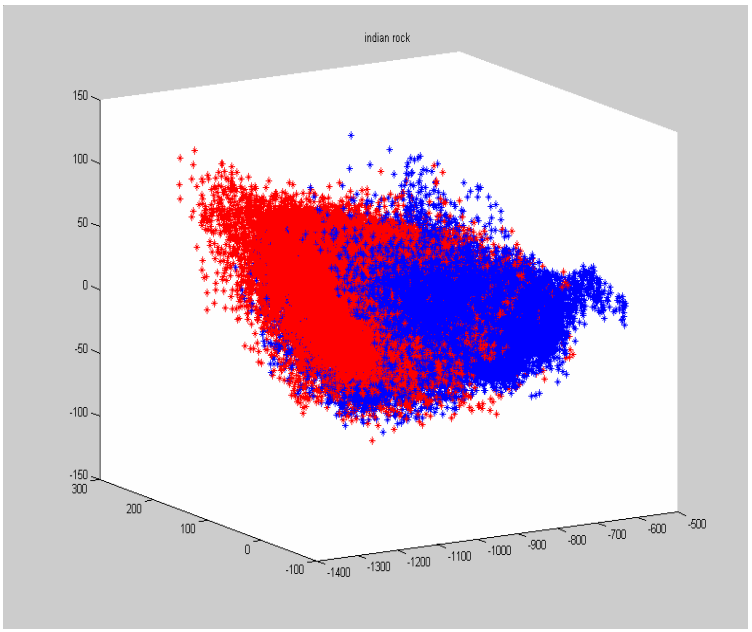

Figure7: Ambient v/s Rock

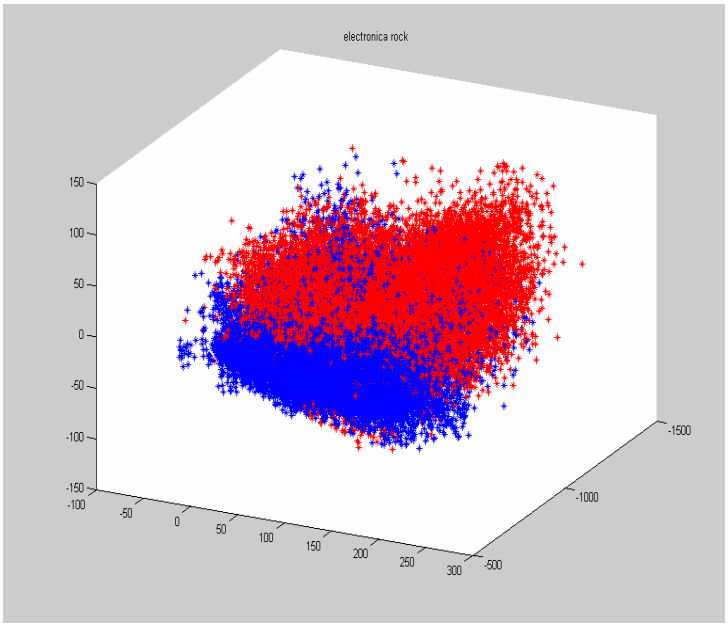

Figure6: Electronic v/s Rock

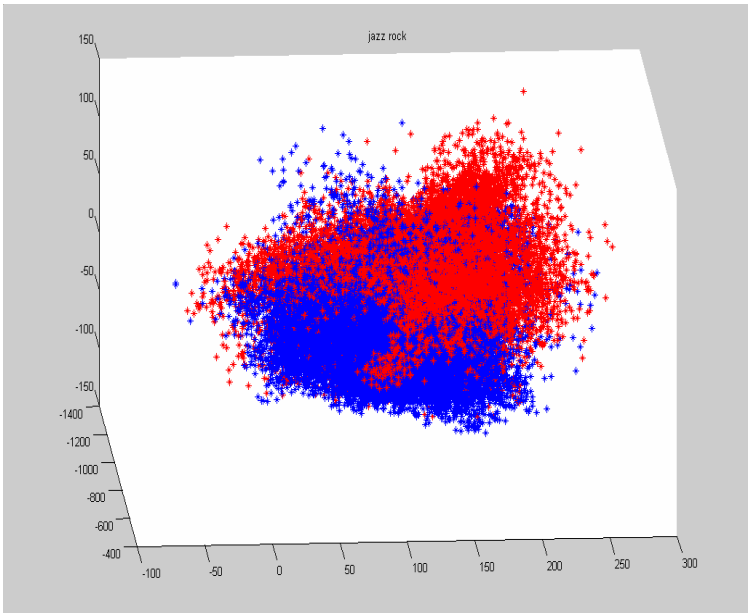

Figure8: Jazz v/s Rock

2.2 Classification

Second step in building a robust classification system is to input the feature values extracted above into a classifier which operates a set of rules to generate various codebooks. In this project we have used Vector Quantization (VQ) as a classifier, which is the most widely used classification technique in the field of speech and audio recognition/classifier. Vector quantization is basically a concatenation of Binary Split and K-Means algorithms.

From feature extraction step, we extract two sets of feature vectors for each genre respectively. These feature vectors are saved in files named ***genre.mfcc*** and ***genre.energy*** for all the 9 genres. The dimension of these feature vector files are generally of the order 43000 x 20 for *genre.mfcc* and 43000 x 7 for *genre.energy*. It is not recommended computationally to compare an input test feature vector with all 43000 feature vectors. Hence we use Vector Quantization to extract 128 codebook vectors corresponding to each feature vector file. Finally we have two codebooks corresponding to each genre. These codebook vectors are saved in two files named ***genre.mfcc.codebook*** and ***genre.energy.codebook.*** The size of these codebooks is 128 x 20 for genre.mfcc.codebook and 128 x 7 for genre.energy.codebook.

In the testing phase, we extract the same sets of feature vector for the input audio files. For a 30 second test audio file, we generally get about 2000 mfcc feature vector and 2000 energy feature vector. We finally compute the distance of these feature vectors from the codebooks saved before. Depending upon the minimum distance measure, we finally annotate the input audio file with the corresponding genre name.

2.2.1 Results

We developed a database comprising of about 2000 audio files from 9 different genres. These audio files are downloaded from the website *www.magnatune.com* [22], which has been a source of audio database for audio processing over the past decade. We have used roughly 120 audio file per genre of 30 second each (excluding initial 30 seconds of the audio file) during the training phase. This accounts to almost 60 minutes of training dataset. While testing we have used about 80 audio files per genre. This accounts to about 40 minutes of testing dataset.

Here are the results obtained after the exhaustive training and testing phases:

	Ambient	**Blues**	**Classical**	**Electronic**	**Folk**	**Jazz**	**New Age**	**Rap**	**Rock**
Ambient	**64**	0	12	6	6	6	4	2	0
Blues	0	**76**	0	6	2	6	0	0	10
Classical	0	0	**96**	0	0	0	2	2	0
Electronic	0	0	2	**94**	0	2	2	0	0
Folk	0	2	6	0	**88**	2	2	0	0
Jazz	0	0	0	6	0	**88**	0	6	0
New Age	2	3	15	0	3	10	**67**	0	0
Rap	0	1	0	5	0	0	0	**90**	4
Rock	0	0	0	2	2	5	0	6	**85**

Figure9: Confusion Matrix using MFCC's as feature vector.

	Ambient	**Blues**	**Classical**	**Electronic**	**Folk**	**Jazz**	**New Age**	**Rap**	**Rock**
Ambient	**74**	0	10	4	4	4	4	0	0
Blues	0	**80**	0	4	0	6	0	0	10
Classical	0	0	**96**	0	0	0	0	4	10
Electronic	0	0	0	**96**	0	0	4	0	0
Folk	0	2	4	0	**92**	0	2	0	0
Jazz	0	0	0	4	0	**90**	0	6	0
New Age	3	0	13	0	2	7	**75**	0	0
Rap	0	2	0	5	0	0	0	**90**	3
Rock	0	0	0	5	0	1	0	6	**88**

Figure10: Confusion Matrix using Log-Compressed Energies as feature vector.

3 Conclusion

The obtained classification accuracy of 86.77 % and 83.11 % is encouraging and comparable to other state-of-the-art algorithms evaluated at the MIREX contest (annual Audio Processing algorithm contest). By analyzing carefully confusion matrices, one can notice that classification errors make sense: for example, on the MAGNATUNE dataset, 20.59 % of Punk excerpts are classified as Rock. There is indeed a clear overlap between these two genres and the misclassified examples may have been probably better described as belonging to both classes.

From the above confusion matrix it is also evident that Log-Compressed energies give better representation of the audio signal as compared to MFCC's feature vectors. Using Log-Compressed energy we were able to achieve 86.77 % of accuracy as compared to 83.11 % of classification using MFCC's as feature vectors. Overall we suggest using both feature vectors for audio genre classification task, as MFCC's give a representation of musical timbre which is equally important as Log-compressed energies which gives better representation of perception model for audio files.

Overall, we find that research is evolving from purely objective machine calculations to techniques where learning phases, training data sets, and preliminary knowledge strongly influence performance and results. This is particularly comprehensible for music genre classification, which has always been influenced by experience, background and sometimes personal feeling. But even in several other classification domains, music related or not, many outstanding solutions exist where machine learning plays a fundamental role, complementary to signal processing.

4 Future Directions

In the first phase of the project we have mainly focused on implementing the most basic feature vectors and classification techniques that have been used world wide in development of a robust Automatic genre classifier. For the coming phase, we would like to focus on other feature vectors which extract the exact signal properties from the given audio signal. Use of better classification schemes like GMM, SVM and HMM can also be a direction of research in the coming semester, as they have given better classification results in the field of speech and audio classification systems world wide.

We would also like to integrate our algorithms and techniques in building up an Automatic Playlist generation system. Every listener has his own choice for music, and he wants to have access to similar sort of music from a given database. Automatic Playlist generator system will take an input query song from the user and help him achieve a Playlist of songs, where songs will be musically identical to his input query song.

5 References

[1] R. Dannenberg, J. Foote, G. Tzanetakis, and C. Weare, “Panel: new directions in music information retrieval,” *Proc. Int. Computer Music Conf*., Habana, Cuba, Sept. 2001.

[2] F. Pachet and D. Cazaly, “A taxonomy of musical genres,” *Proc. Content-Based Multimedia Information Access (RIAO)*, Paris, France, 2000.

[3] F. Pachet, J.J. Aucouturier, A. La Burthe, A. Zils, and A. Beurive, “The cuidado music browser: an end-to-end electronic music distribution system,” *Multimedia Tools Applicat.*, 2004, Special Issue on the CBMI03 Conference, Rennes, France, 2003.

[4] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” CUIDADO I.S.T. Project Rep., 2004.

[5] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[6] A. Meng, P. Ahrendt, and J. Larsen, “Improving music genre classification by short-time feature integration,” *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005, pp. 604–609.

[7] K. West and S. Cox, “Finding an optimal segmentationfor audio genre classification,” *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005, pp. 680–685.

[8] N. Scaringella and G. Zoia, “On the modeling of time information for automatic genre recognition systems in audio signals,” *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005, pp. 666–671.

[9] E. Gomez, A. Klapuri, and B. Meudic, “Melody description and extraction in the context of music content processing,” *J. New Music Res.*, vol. 32 no. 1, 2003.

[10] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003

[11] A. Berenzweig, D. Ellis, and S. Lawrence, “Using voice segments to improve artist classification of music,” *Proc. AES 22nd Int. Conf. Virtual, Synthetic Entertainment Audio*, 2002.

[12] M. Mandel and D. Ellis, “Song-level features and support vector machines for music classification,” *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005, pp. 594–599.

[13] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, “Recognition of music types,” *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*, Seattle, WA, USA, 1998, vol. II, pp. 1137–1140

[14] N. Casagrande, D. Eck, and B. Kegl, “Geometry in sound: a speech/music audio classifier inspired by an image classifier,” *Proc. Int. Computer Music Conf. (ICMC)*, 2005.

[15] A. Flexer, E. Pampalk, and G. Widmer, “Novelty detection based on spectral similarity of songs,” *Proc. 6th Int. Symp. Music Information Retrieval*, London, UK, 2005, pp. 260–263.

[16] K. West, S. Cox, “Features and classifier for the automatic classification of musical audio signals”, *Proc. Of the 5th Int. Conf. on Music Information Retrieval (ISMIR), Barcelona, Spain*, 2004.

[17] G. Tzanetakis, P. Cook, “Musical genre classification of audio signals”, *IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5*, July 2002.

[18] N. Scaringella, G. Zoia, “On the modelling of time-information for automatic genre recognition systems in audio signals”, *Proc. of the 6th Int. Conf, on Music Information Retrieval, London, UK*, 2005.

[19] L. Rabiner, B.H. Juang, *Fundamentals of speech recognition*, Englewood Cliffs, NJ, Prentice-Hall, 1993.

[20] J.J. Aucouturier, F. Pachet, “Improving timbre similarity: how high’s the sky?”, *Journal of Negative Results in Speech and Audio Sciences*, 2004.

[21] http://www.music-ir.org/

[22] http://www.magnatune.com