Song genre classification

Pakshal Bohra, Himanshu Pandotra, Anand Pathak, Rishabh Iyer

***Abstract*—Analyzing music audio files based on genres and other qualitative tags is an active field of research in machine learning. Today, people using online music services are very likely to search for music by genre, so understanding how to automatically classify music by genre is growing in importance. However music genre classification has been a challenging task in the field of music information retrieval (MIR). Music genres are hard to systematically and consistently describe due to their inherent subjective nature. Any such problem comprises two steps, the first being a pre-processing stage that primarily revolves around feature extraction and the second being the usage of machine learning classifiers. In this paper, we have relied purely on Mel Frequency Cepstral Coefficients (MFCC) to characterize our data and then applied machine learning techniques such as k-nearest neighbor (kNN),k-means, multi-class SVM, and neural networks to classify the following four genres: classical, jazz, metal and pop. Our classifiers achieve an accuracy of upto 55 %**

I. Introduction

Given the explosion of multimedia data on the internet today the identification and classification of musical data is an important and growing problem. With most musicians/ artists associating themselves with a particular genre, genre based searches have risen substantially in recent years. Music genre is also a key factor when it comes to online music recommendation systems. Till date, genre classification is primarily done manually by appending the genre data as metadata along with audio files or including it in the album information. Our work aims to create an automatic content-based classification mechanism that utilizes information within the audio rather to classify a song into a particular genre.

II. Challenges

1) **Defining a genre:** Unfortunately a genre is not a very well defined term. In the past, genres have been based on either the progenitor of a type of music, the time period when the time of music flourished or even based on geographical locations where it has flourished. Also several genres do not have well defined boundaries, i.e. more often than not songs overlap these boundaries and genre identification depends on the listener.
2) **Non-stationarity of speech signals:** It is observed that the speech spectrum is highly variable in the sense that it contains a lot of overlapping frequency components which are continuously varying in time. This gives rise to non-stationarity in the speech signal. General signal processing techniques like the Fourier transform make an underlying assumption on the stationarity of the signal whose spectrum is to be computed. Hence to find spectra of speech signals, we need to consider alternative techniques.
3) **Defining window size:** To overcome the above problem, it is assumed that speech remains stationary in short frames of time. Hence we divide the given speech signal into several frames of very short time intervals and then apply the normal signal processing techniques. The parameters such as frame size to be taken depend on the type of signal, and is especially more difficult to determine in case of music signals which demonstrate very high non-stationarity.
4) **Particular problems with music files:** The problem of music signal processing is more difficult compared to speech processing. The frequencies of resonances of the vocal tract(formants) are governed primarily by the constrictions in mouth and shape of vocal tract. Since the shape of the vocal tract cannot change significantly, the normal speech signal spectrum can be predicted to a certain extent. However, in case of music signals, there is no consistent, well defined long term spectrum.Moreover, there exists variability across speakers and different types of music.

III. Dataset

We have used the GTZAN dataset from the MARYSAS website[3]. This is a curated dataset, i.e. humans assign their tags selectively with accuracy in mind for academic purpose. It contains 9 music genres, each genre has 100 audio clips in .au format. The genres are blues, classical, country, disco, pop, jazz, reggae, rock, metal. Each audio clip is a 22050Hz Mono 16-bit file and is 30s long. We imported these files using the

Advisor: Prof. Ganesh Ramakrishnan, IIT Bombay

au_read function from the sunau module in python. We also split the dataset in a 70:30 ratio for training and test purposes. We have identified that the genres of classical, jazz, metal and pop are symbolic of typical problems faced in song genre classification. Metal and pop are difficult to separate from one another and the same goes for jazz and classical music. These two pairs however are significantly different from one another. Hence we have chosen to work with only these 4 genres.

IV. Feature Extraction

We are using the Mel-frequency cepstral coefficients (MFCC) to characterize each song.

A. What are MFCCs ?

MFCCs are a set of coefficients which collectively represent short term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency.

B. Why MFCCs?

1) There are several features which can be used for classification purpose like Linear prediction coefficients(LPCs), short time energy, short time autocorrelation, etc.
2) However, MFCCs are preferred over these as they are better able to model the natural speech production and perception as is done by the vocal cords and the ear respectively.
3) As an example, the LP coefficients use an all-pole model for speech production, hence the zeroes in the transfer functions of sounds such as nasal sounds cannot be modelled efficiently.
4) The properties of sound perceived by the ear do not bear linear relationship with the actual properties. Hence certain amount of processing needs to be done to model these effects of perception in speech processing, which is accounted for by the MFCC coefficients.

C. Derivation of MFCCs:

Derivation of MFCCs from a given sound signal is done in the following steps:

1) **Frame signals into short frames:** – to overcome the non-stationarity of sound signals. Tradeoff-short frames-not enough samples for reliable spectral estimate, large frames- bring in non-stationarity
2) **For each frame calculate the periodogram estimate of the power spectrum:** – The basilar membrane(BM) in the human ear shows high vibration for a given frequency at a specific distance from the base. This is analogous to power spectral estimate which which estimates the power contained in different frequency ranges.
3) **Apply mel filterbank to power spectra:** Resolution of ear to frequencies decreases with increasing freqency. Hence, we use filter bank with bandwidths of filters increasing with frequency(triangular filters).
4) **Take logarithm of all filterbank frequencies:**Loudness is perceived by the human ear on a logarithmic scale.
5) **Take DCT of log filterbank energies:** DCT decorrelates the energies which are correlated due to the overlapping filter banks.
6) **Keep the top 13 DCT coefficients:** They contain most of the information in the sound signals.

We have followed two separate approaches to using the MFCC coefficients as a feature vector. In the first, as recommended by [4], we further reduce this matrix representation of each song by taking the mean vector and covariance matrix of the cepstral features over each 20ms frame, and storing them as a cell matrix. This is equivalent to modelling the frequencies as a multivarate gaussian distribution and greatly reduces further computational requirements. We then combine the mean vector and the top half of the covariance matrix (since it is symmetric) into one feature vector. As a result, we get $15 + (1 + 15) * (7.5)$ features for each song.
Another method that we explored to utilize the MFCC coefficients as a feature space was using the bag of tones model. The bag of tones model derives its inspiration from the oft-used bag of words model. In this model, a text is represented as an unordered collection of words, disregarding grammar and even word order. The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier. To create the bag of tones model used in [1] we do the following. First given that we now have 13 MFCC coefficients for each window in each song, we represent each frame as a point in 13-dimensional space. We then run a K-means clustering on this set of points. Now for each song, we calculate the cluster that each of its frames belong to and plot a histogram where the number of bins are equal to the number of clusters or K. This histogram is then treated as the feature vector and given as an input to the various classifiers.

V. Classifiers used

Once the feature vector for each song has been created, we treat this problem as a multi-class classification

problem. We have explored the following classifiers for the same.

A. Neural Networks

B. SVM

C. K-Nearest Neighbours

D. Random Forest

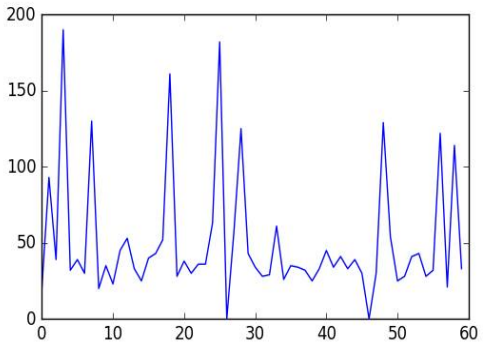

Figure 1: Sample feature vector for classical

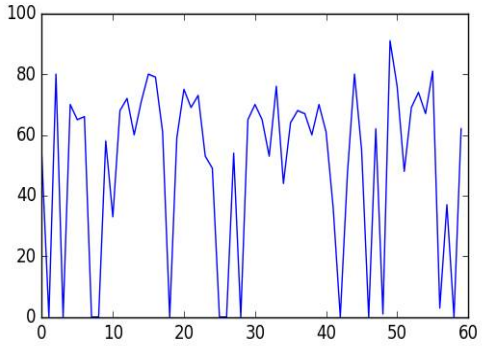

Figure 2: Sample feature vector for Jazz

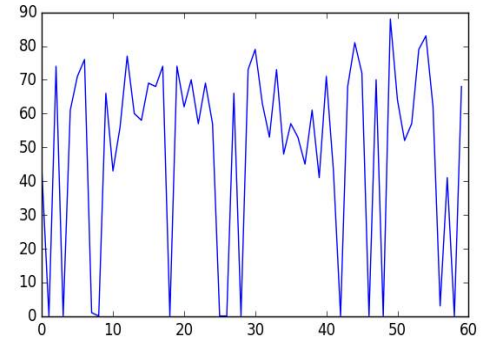

Figure 3: Sample feature vector for Metal

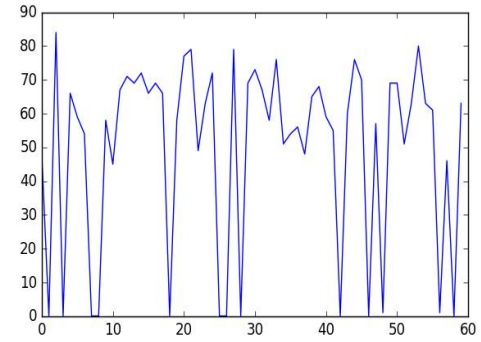

Figure 4: Sample feature vector for Pop

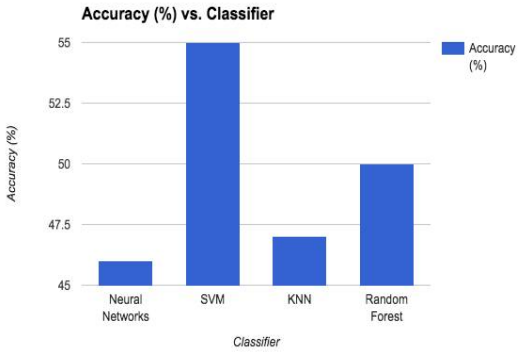

Figure 5: Accuracy across classifiers

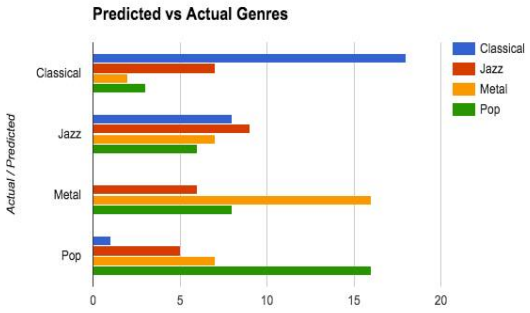

Figure 6: Predicted values vs. actual values of genres for SVM

Our prediction saturates around 55 % . This is due to the high pairwise correlation between the genres of metal and pop and the genres of jazz and classical music. The prediction for jazz is especially poor.

VI. What we learnt

- The challenges faced in song genre classification
- Extracting MFCC vectors which can be used as features in most of the sound processing related applications.

- Learnt about the bag of tones model which finds its way in several applications of machine learning.
- The behaviour of different classifiers for the song genre detection problem.

VII. Conclusion

Music genre classification is an open problem and the major problem in this classification problem is the extraction of relevant features. We have used the Mel Frequency Cepstral Coefficients in a variety of ways to try and characterize the genre but we believe that the feature space needs to be improved further. Our classifiers achieve an accuracy of upto 55 % in attempting to classify across four main genres. Future work includes adding more features that can better represent the song genre. Also given the probability of overlap between genres, usage of Gaussian mixture models is a possibility.

References

[1] Qin et. al A Bag-of-Tones Model with MFCC Features for Musical Genre Classification
[2] Diab et.al Musical Genre Tag Classification With Curated and Crowdsourced Datasets
[3] Tzanetakis, George, Georg Essl, and Perry Cook. ”Automatic Musical Genre Classification Of Audio Signals.” The International Society for Music Information Retrieval. 2001.
[4] Haggblade et.al Music Genre classification