Music Genre Classification

Khursheed Ali (163059009)

Anshul Gupta (16305R001)

Abhijeet Dubey (16305R006)

Nithin S (16305R007)

Introduction

- Sometimes it happens that we listen to a particular music, we instantly develop an affinity towards that genre and want to listen to same type of music.
- Or sometimes we just want to organize our music collection based on genre.This project aims to classify music into different categories such as:
  - Blues
  - Classical
  - Country
  - Disco
  - Reggae
  - Hip Hop
  - Jazz
  - Metal
  - Pop
  - Rock

Feature Extraction

- Focusing on Timbral Features.
- The features used to represent timbral texture are based on standard features proposed for music-speech discrimination and speech recognition.
- Based on the short time Fourier transform (STFT) and are calculated for every short-time frame of sound.

Spectral Centroid

- Defined as the center of gravity of the magnitude spectrum of the STFT

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}$$

Where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n.

- The centroid is a measure of spectral shape and higher centroid values correspond to “brighter” textures with more high frequencies.

Feature Extraction

- The time duration for each audio file is slightly different.
- With a sampling rate of 22050 we get ~660000 samples / file.
- We are finding centroid for each batch of sample.
- Size of each batch being 2048 and hop length of 512, thus getting approximately 1290 centroids per audio file.
- same process for other features.

```
--------------------[reggae]-----------------------
('reggae', ' ymin:', 661504, ' max:', 661794)
--------------------[hiphop]-----------------------
('hiphop', ' ymin:', 660000, ' max:', 675808)
--------------------[classical]--------------------
('classical', ' ymin:', 661344, ' max:', 672282)
--------------------[rock]-------------------------
('rock', ' ymin:', 661408, ' max:', 670340)
--------------------[country]----------------------
('country', ' ymin:', 661100, ' max:', 669680)
--------------------[blues]------------------------
('blues', ' ymin:', 661794, ' max:', 661794)
--------------------[jazz]-------------------------
('jazz', ' ymin:', 661676, ' max:', 672100)
--------------------[metal]------------------------
('metal', ' ymin:', 661504, ' max:', 661794)
--------------------[pop]--------------------------
('pop', ' ymin:', 661504, ' max:', 661504)
--------------------[disco]------------------------
('disco', ' ymin:', 661344, ' max:', 668140)
```

Spectral Rolloff

- The spectral rolloff is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n].$$

The rolloff is another measure of spectral shape.

Time Domain Zero Crossings

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|$$

where the *sign* function is 1 for positive arguments and 0 for negative arguments and *x[n]* is the time domain signal for frame *t*. Time domain zero crossings provide a measure of the noisiness of the signal.

Mel-Frequency Cepstral Coefficients

- Perceptually motivated features that are also based on the STFT.
- After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling.
- Finally, in order to decorrelate the resulting feature vectors a discrete cosine transform is performed.
- Although typically 13 coefficients are used for speech representation, we have found that the first five coefficients provide the best genre classification performance.

Other Features

- Root Mean Square Energy.
- Spectral Contrast: considers the spectral peak, spectral valley and their difference in each sub-band.
- Overall, the feature vector for describing timbral texture consists of the following features: means and variances of spectral centroid, rolloff, zerocrossings over the texture window , rms energy, spectral contrast, and means and variances of the first five MFCC coefficients over the texture window.
- Results in a 20-dimensional feature vector.

Librosa

- The features were extracted using the python package LibROSA.
- LibROSA is a python package for music and audio analysis.
- It provides the building blocks necessary to create music information retrieval systems.

Approaches using in-built libraries

- Once the feature extraction is done and we have the dataframe, we can apply standard off the shelf python libraries for classification.

	Trained Misclassified Points	Trained Accuracy	Test Misclassified Points	Test Accuracy
Decision Tree	215	73	95	52
Random Forest	63	92	80	60
K-Neighbors	401	49	145	27
Gradient Boosting	1	99	85	57
Logistic Regression	385	51	105	47
Support Vector	631	21	156	22
Bernoulli NB	638	20	154	23
Gaussian NB	462	42	113	43
Adaboost	539	32	146	27

Approaches using our implementations

- The dataset is split into 80:20 ratio between training dataset and test dataset randomly.
- The dataset then remains fixed for all training models.

Multi-class Neural Network

- Hidden Layers : 7
- Neurons / Hidden Layer: 40
- Learning Rate : 0.01
- Iterations : 1000

	Misclassified	Accuracy
Train	10	98.0 %
Test	112	44.0 %

Decision Trees

- Maximum depth of tree : 20
- Minimum record count : 2

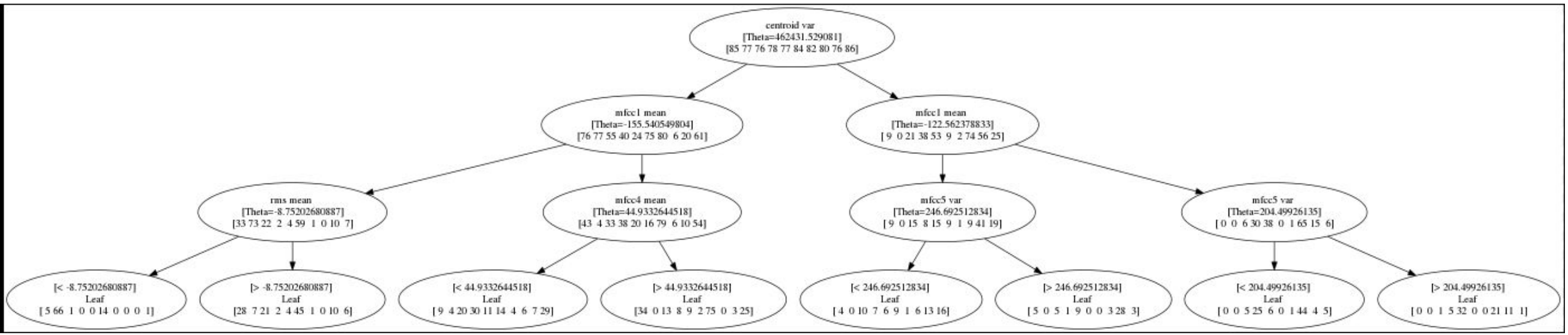
	Misclassified	Accuracy
Train	0	100.0 %
Test	100	50.0 %

- Outperforms 6 out of 9 built in library functions.

Decision Trees

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	8	0	4	0	0	2	0	0	1	0
classical	1	19	0	0	0	2	0	0	0	1
country	3	1	8	3	3	2	0	0	4	0
disco	1	1	3	11	0	1	0	2	1	2
hiphop	1	0	0	1	11	0	2	4	3	1
jazz	1	1	3	0	0	12	0	0	0	0
metal	4	0	0	0	0	0	13	0	0	1
pop	0	0	1	4	3	1	0	8	3	0
reggae	0	0	2	4	5	0	0	3	10	0
rock	3	0	3	4	2	0	2	0	0	0

Decision Trees



Random Forest

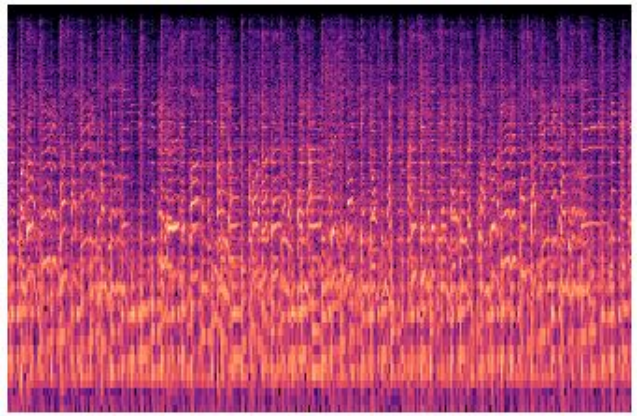
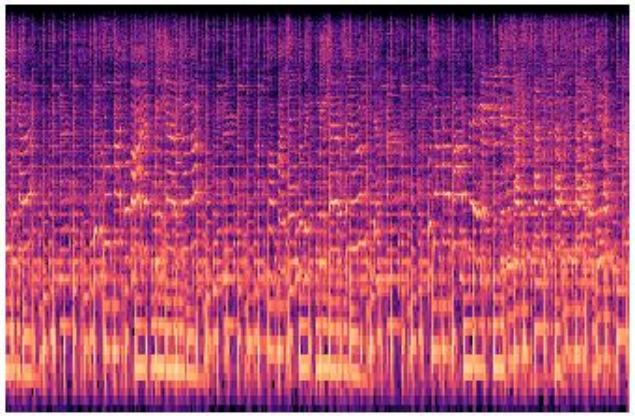
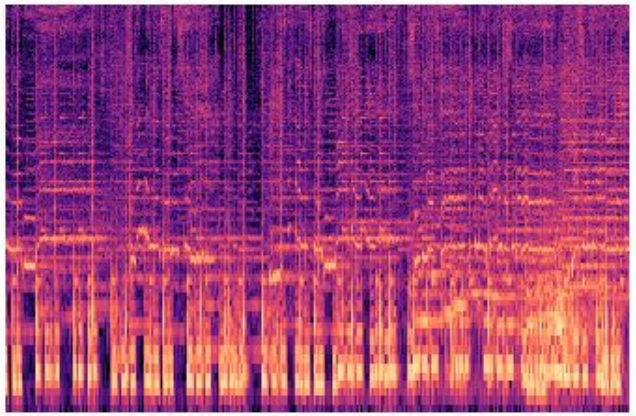
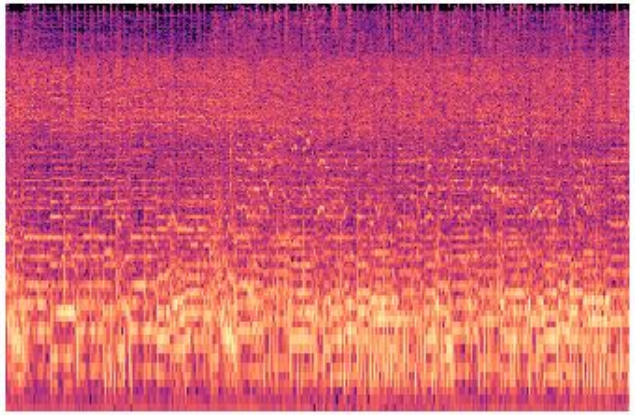
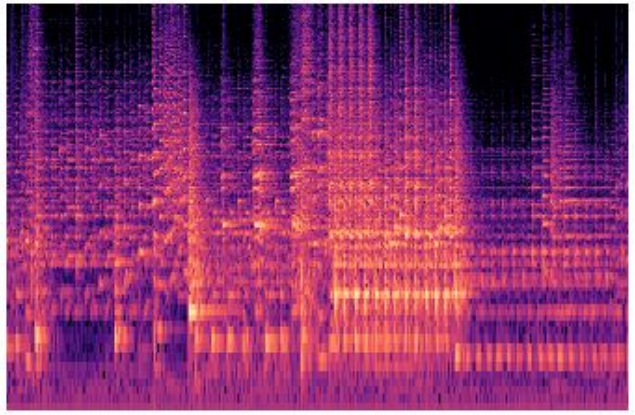
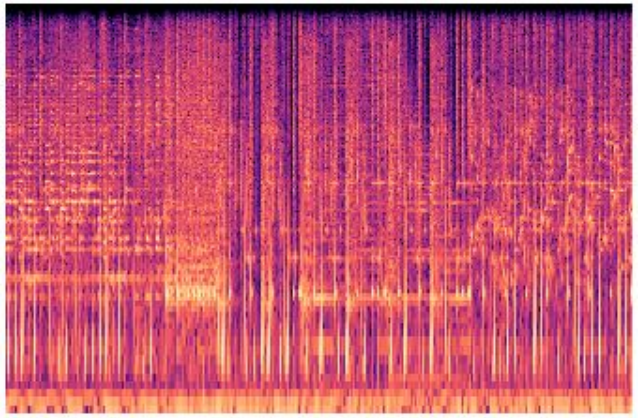
- Wrapper around the decision tree mode.
- We randomly partition the dataset into multiple data subsets and train the decision tree on each subset.
- Number of Trees : 50
- Max. Depth of Tree : 7
- Min. Record Count : 20

	Misclassified	Accuracy
Train	8	99.0 %
Test	73	63.5 %

Convolutional Neural Network

- Attempt at classification of genres using the Power Spectrogram of the audio files.
- We processed the audio files and came up with the spectrograms.
- The spectrograms are quite distinct from each other.
- Each genre has a particular rhythm which is captured in this image.
- We tried to classify the genres by training a Convolutional neural network on these images.
- We used tensorflow library for creating the model.

Convolutional Neural Network

Convolutional Neural Network

- Image size is 200 x 200 x 4
- 2 pairs of Convolutional and Max Pooling Layers
- First convolution layer has 30 features
- Second convolution layer has 15 features
- A fully connected hidden layer of size 128
- A fully connected output layer of size 10
- Batch Gradient Descent learning rate is 1e −4 and batch size of 50
- Since the data set used is very small for a neural network to be fully trained, the accuracy on test data was **0.228924**

Observations and Conclusions

- The original paper achieved the accuracy of 61% using Timbral Features as well as Rhythmic Features of the audio files.
- We were able to achieve more accuracy (63.5 %) than them using only Timbral Features.
- Also we were able to achieve much better accuracy than most standard library implementations of classification algorithms.

TimeLine

- Before Mid stage:
  - Analysis of spectrum of audio files
  - Using standard neural network library implementation for classification
- After Mid Stage:
  - Setting up audio processing libraries
  - Pre-processing of audio files
  - Classification using standard off-the-shelf libraries
  - Building our own models
    - Multi-Class Neural Network
    - Decision Tree
    - Random Forest
    - Convolutional Neural Network

Contribution

- Khursheed Ali : Decision Trees, Random Forest
- Anshul Gupta : Convolutional Neural Network, Standard Libraries(KNN, Multinomial, Bayesian,SVC)
- Abhijeet Dubey : Preprocessing, spectral analysis
- Nithin S : Multi-Class Neural Networks

- Feature Extraction, Report, Presentation: Everyone Contributed

Future Work

- We are only using Timbral features and still able to achieve a good accuracy.
- if we also include Rhythmic features in our dataset, we may be able to achieve much higher accuracy then what we are getting now.
- Dataset which we used in this project was very small (only 1000 audio files) and not sufficient to train a neural network.
- We can train our models on a much bigger dataset. (A million song dataset is added in reference)

References

- [1] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classifi-cation. InProceedings of the 26th Annual International ACM SIGIR Conference on Research andDevelopment in Informaion Retrieval, SIGIR ’03, pages 282–289, New York, NY, USA, 2003. ACM.
- [2] A¨aron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recom-mendation. InProceedings of the 26th International Conference on Neural Information ProcessingSystems, NIPS’13, pages 2643–2651, USA, 2013. Curran Associates Inc.
- [3] Helge Homburg, Ingo Mierswa, B¨ulent M¨oller, Katharina Morik, and Michael Wurst. A benchmarkdataset for audio classification and clustering. InProceedings of the 6th International Conference onMusic Information Retrieval, London, UK, September 11-15 2005.http://ismir2005.ismir.net/proceedings/2117.pdf.
- [4] G. Tzanetakis and P. Cook. Gtzan genre collection.http://marsyas.info/downloads/datasets.html.
- [5] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, andOriol Nieto. librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and JamesBergstra, editors,Proceedings of the 14th Python in Science Conference, pages 18 – 25, 2015.
- [6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million songdataset. InProceedings of the 12th International Conference on Music Information Retrieval (ISMIR2011), 2011