Automatic Speech Recognition - CS753

Music Genre classification and recommendation system

-By

Anjali Nainani (17D110024)

Sankalp Mule (17D110007)

Tushar Mishra (170010027)

INDEX

|  |  |
| --- | --- |
| TOPIC | PAGE NO. |
| Introduction | 3 |
| Dataset & Feature extraction | 4 |
| Model Architecture | 7 |
| Results | 8 |

INTRODUCTION

There has been a deep research for the application of genre classification. Genre classification is the process of grouping the objects together based on similarities. It has been beneficial to many companies, for example Spotify. Companies like Spotify not only focused on classification of songs according to genre but also on recommendation systems for the songs. A recommendation system seeks to predict the "rating" or "preference" a user would give to an item.

In this project our major idea was to build a genre classification model and then a recommendation system based on them. For the genre classification we decided to classify Hindi songs based on different genres. Since no dataset of hindi songs was available, we created our own dataset of genres, Bolly-Sufi, Bolly-pop, Ghazal, Bolly- sad/soft-songs and Carnatic music. We used a CNN model with 5 layers for the task. We tried to build a dataset for the Recommendation system as well but could not make a proper dataset due to time constraints. As a result, we were unable to make a recommendation system.

Dataset & Feature extraction

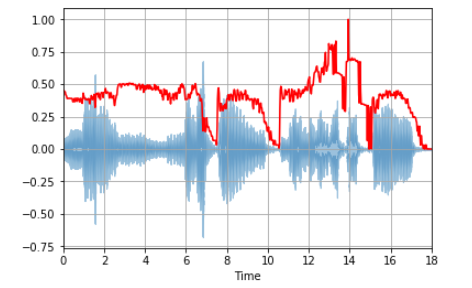
We created a dataset of hindi songs of 5 genres (Bolly-Sufi, Bolly-pop, Ghazal, Bolly- sad/soft-songs and Carnatic music). We used the library ‘librosa’ for pre-processing of the dataset. Using it we restricted the audio size to 30s and then used it for feature extraction

|  |  |  |
| --- | --- | --- |
| Genre | Number of files | Instances from dataset |
| Bolly-sufi | 100 |  |
| Bolly- pop | 100 |  |
| Ghazal | 100 |  |
| Bolly-sad/soft-songs | 100 |  |
| Carnatic music | 100 |  |

Next task was to extract features from the dataset. We used following methods to do so-

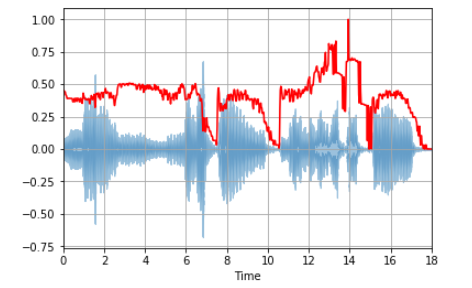
1. **Spectral Roll off**

It is a measure of the shape of the signal. It represents the frequency below which a specified percentage of the total spectral energy, e.g., 85%, lies.



1. **Spectral centroid**

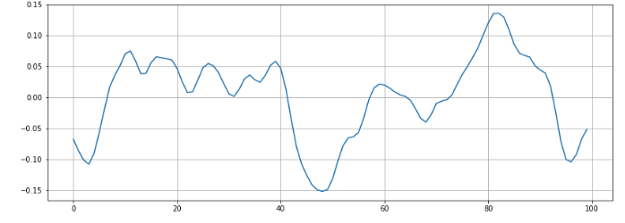
It indicates where the ” center of mass” for a sound is located and is calculated as the weighted mean of the frequencies present in the sound.



There is a rise in the spectral centroid towards the end.

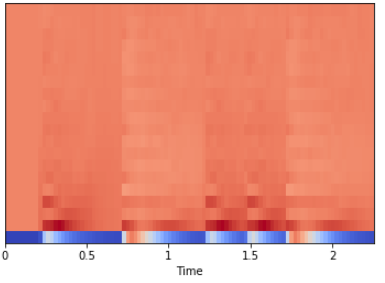
1. **Zero- crossing rate**

It is the rate of sign-changes along with a signal, i.e., the rate at which the signal changes from positive to negative or back.



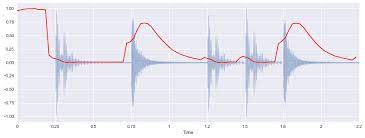
1. **Mel frequency cepstral coefficients (MFCCs)**

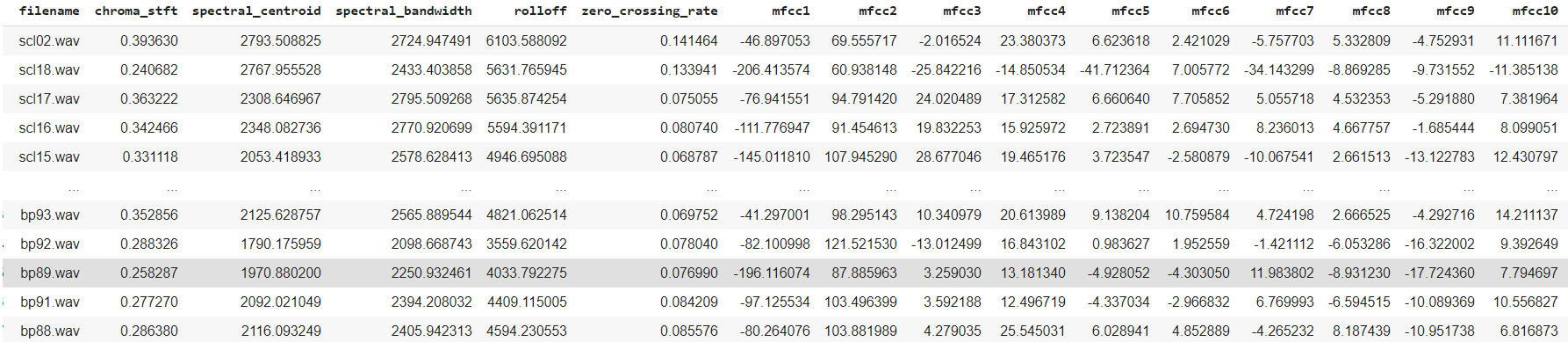
The Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10–20) that concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice.



1. **Spectral Bandwidth**

The spectral bandwidth is defined as the extent of the power transfer function around the center frequency.

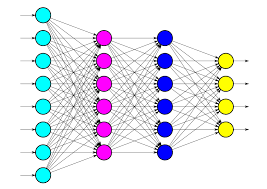




Features extracted for the project

|  |  |  |  |
| --- | --- | --- | --- |
| Layers | Input variables | Output features | Activation used |
| Dense layer1 | (128 x 4522) | (128 x 256) | Relu |
| Dense layer2 | (128 x 256) | (128x 128) | Relu |
| Dense layer3 | (128 x 128) | (128 x 64) | Relu |
| Dense layer4 | (128 x 64) | (128 x 5) | Softmax |

Model Architecture with LSTM



Other parameters used-

|  |  |
| --- | --- |
| Batch size | 128 |
| Epochs | 20 |
| Loss function | Sparse categorical cross-entropy |
| Optimization function | Adam |
| Learning rate | 0.001 |

Results

Initially we used a train set of 400 examples and a validation set of 100 examples and we used the validation set for hyperparameter tuning.

Accuracy obtained on validation set without cross-validation - 80%

We further split or training dataset of 400 examples into 4 groups of 100 each and used 4-fold cross validation for hyperparameter tuning and then tested our model on the test set which contained 100 examples (with 20 examples from each class).

Accuracy obtained on test set with cross-validation – 70%

Confusion Matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Bolly Sufi | Bolly pop | Ghazal | Sad/Soft songs | Carnatic Music |
| Bolly Sufi | 14 | 1 | 3 | 2 | 0 |
| Bolly pop | 2 | 15 | 1 | 0 | 2 |
| Ghazal | 4 | 1 | 14 | 1 | 0 |
| Sad/Soft songs | 1 | 1 | 2 | 16 | 0 |
| Carnatic Music | 0 | 1 | 0 | 0 | 19 |

Class wise Accuracy Scores

|  |  |
| --- | --- |
| Bolly Sufi | 70% |
| Bollywood pop | 65% |
| Ghazal | 70% |
| Bollywood sad/soft songs | 70% |
| Carnatic music | 75% |

**Total F1-score: 69.6%**