Machine Learning Group Project
Speech Music Discrimination
The University of Dublin, Trinity College, Ireland, 2022-2023

Prishita Singh
singhp5@tcd.ie
22306048

Pallavit Aggarwal
aggarwpa@tcd.ie
22333721

Gautam Thapar
thaparg@tcd.ie
22322732

Abstract

We present a classifier that classifies audio samples into music or speech. Our classifier uses several features such as zero-crossing rate, spectral centroid, spectral roll-off, MFCC's (Mel Frequency Cepstral Coefficients) and Chroma Features which capture pitch, timbre, energy, signal change and other soft-information from an audio sample (This is explained in the report). We demonstrate the separability of music-speech samples by using 3 different classifiers. This is a binary classification problem and we use SVM, Naive-Bayes and Neural Network models for this.

Introduction

Music Information Retrieval is a growing domain with increased research happening on Sounds and Music of different types. Given our affinity towards music, we wanted to explore it deeply and look at it in a way that's processed by machines. Our approach led us to process music and sounds numerically using different features that have been defined in the late 20th and early 21st century. MFCC and Chroma Features are used among other features to note the pitch and timbre of the sounds respectively. In our problem statement we differentiate between music and speech.In our problem statement, we have a binary classification scenario of "Human-Speech" and "Music". We apply SVM, Naive Bayes and Neural Network models respectively.

Dataset and Features

Our dataset consists of 84 musical audios from various singers extracted from sound-cloud and 112 speeches extracted from different speakers. A total of 196 clips are used to train the model. All the audios are clipped to X seconds and converted to (.wav) format . Stereo files in the dataset are converted to mono waveform. Below is the snapshot of the dataset used:

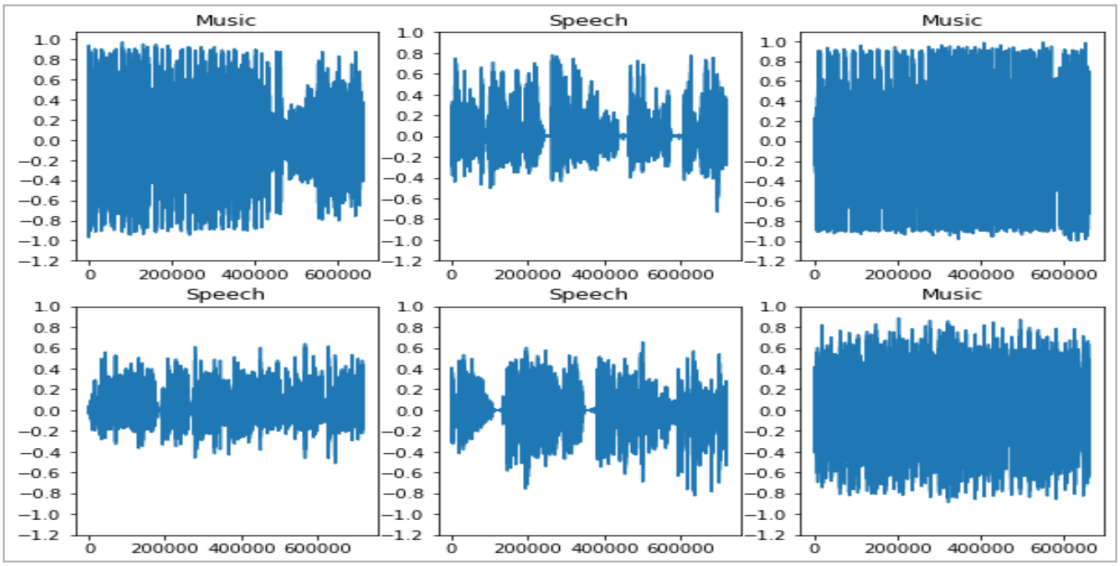
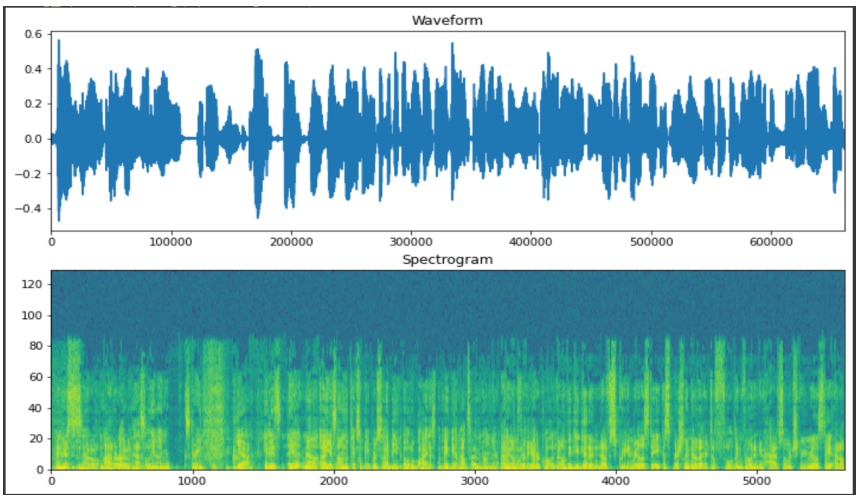


Figure 1: Waveform plotted for the dataset(i.e. audio and speech)

For data augmentation, musical audios and podcasts are picked from various singers and speakers. Moreover, special audios such as rap tracks i.e.speech following a synchronous musical tone, chit chatting noises are added to the model to ensure some noise is added as well. We also included the beats from musical instruments to have high quality data.



The following attributes from the audio are extracted from the audio Files:

1. Mel-Frequency Cepstrum Coefficients:
2. Chroma
3. Zero Crossing Rate
4. Spectral Centroid
5. Spectral Bandwidth
6. RMS of each frame

Figure 2: The Waveform and Spectrogram plotted for a Speech

Methods

CNN Model:

MFCC can be used to represent the audio files into a pictorial format. Thes dataset of these images would be the input in our model and the predictions from these would be our model. The MFCC from the dataset from each clip is passed as the input instead of converting the MFCC data obtained for each audio into an Image and then applying CNN layers to convolve the image and flattening the results to obtain the desired output. A sequential layer model is used to build the model.

Naive Bayes:

Naive Bayes is a statistical classification technique based on Bayes Theorem. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, the MEL Coefficients and Chroma features are computed differently and are not interdependent on each other thereby they attain conditional independence. (On a closer look, these parameters are closely related in the way they're computed, but the assumption of discrete independence works very well for NB). The features are plotted to show continuous data but their discrete values are extracted by taking mean.

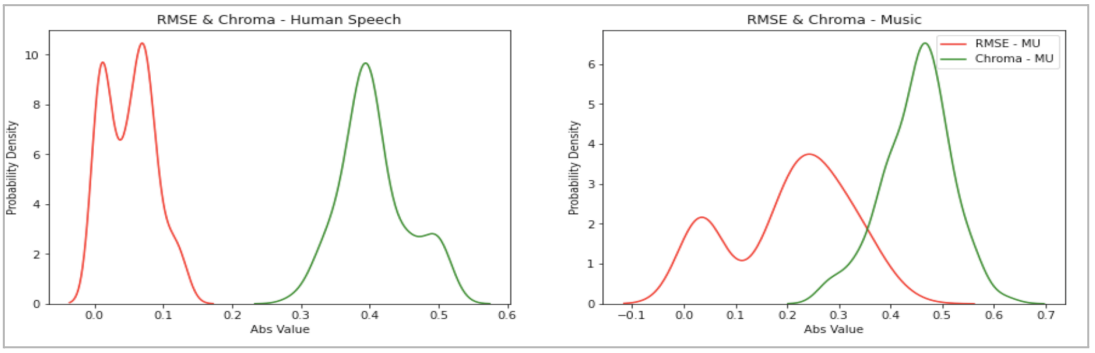


Figure 3: RMSE & Chroma plotted as normal distribution

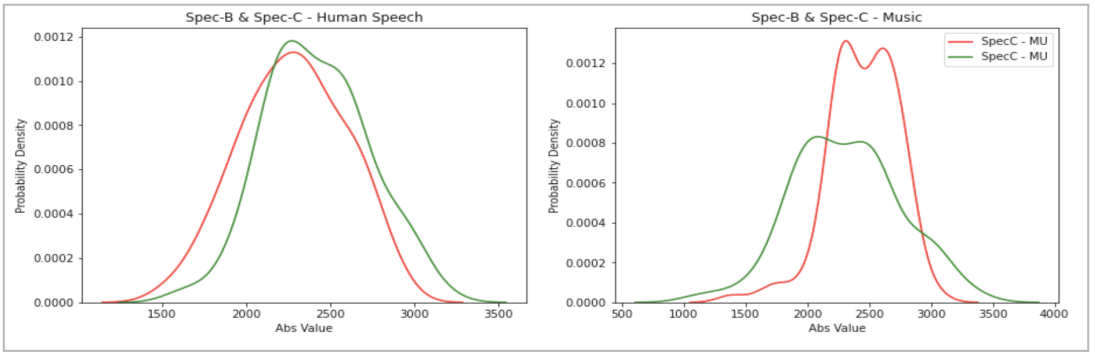


Figure 4: SPEC-BW and SPEC-C plotted as normal distribution

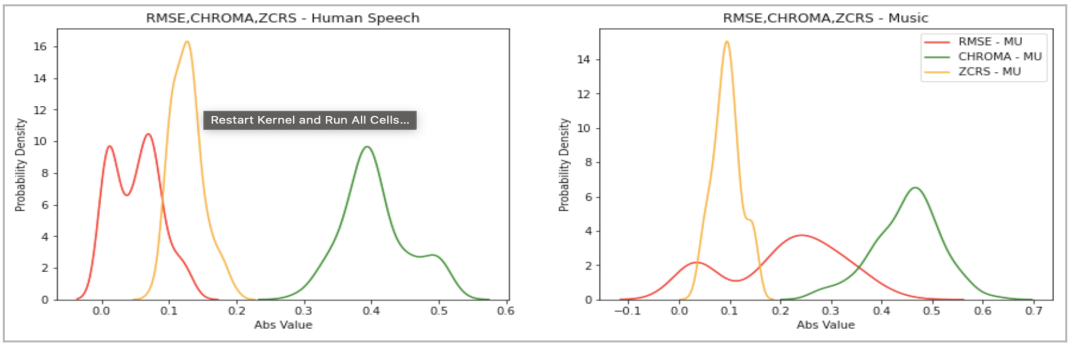


Figure 5: SPEC-BW and SPEC-C plotted as normal distribution

Plotting the features like Chroma, MFCCs, RMSE etc of human speech and music show a normal distribution observed (A gaussian bell curve, though not perfect) over the feature set. Hence application of Gaussian Naive Bayes can be done.

SVM Model: SVM is a supervised machine learning algorithm. It uses classification techniques for two-group classification problems. SVM takes data points and draws a decision boundary also known as a hyperplane. For example, we have two tags music and speech and we have two features MFCC and Chroma. SVM generates the hyperplane using these data points. The best hyperplane is the one that maximises the margins from both tags. SVM works best when there is a clear separation between tags like music and speech. It is also memory efficient. The trade-offs for this model is that it does not work well with large data and it does not give probability estimates directly.

Experiments, Results & Discussion

The following primary metrics are used:

- **Accuracy:** Formula: (True positives + True negatives) / (True positives + True negatives + False positives + False negatives) Representing the correctly classified instances divided by the total number of instances. Comparison of accuracies between the test and training data can identify overfitting in the model efficiently.
- **Precision:** It is the ratio of (True positives) / (True positives + False positives) , representing the correctly classified number of false positives.
- **ROC Curve:** It represents the true positives against the false positives. It is a probability curve.
- **Area Under Curve (AUC):** Area covered under the ROC Curve is called the AUC. It describes how efficiently a model can segregate the classes.
- **Recall:** Formula: It is the ratio of (True positives) / (True positives + False negatives) , representing the correctly classified number of false positives.
- **F1 Score:** It is the harmonic mean of precision and recall and helps to predict the accuracy of binary classifiers.
- **Confusion Matrix:** Summary of correctly and incorrectly predicted classes by the model.

CNN Model:

	feature	class_label
0	[-273.61823, 101.61787, -2.399586, 46.101913, ...	Human-Speech
1	[-228.65521, 106.11402, -15.667508, 50.898582,...	Human-Speech
2	[-272.7577, 79.549225, -8.291422, 46.874435, 6...	Human-Speech
3	[-272.6477, 83.50246, -1.5155042, 48.30416, 11...	Human-Speech
4	[-295.90656, 86.00357, -14.50969, 40.37063, 15...	Human-Speech

Alongside is the data obtained after labelling the audios against a defined Class Label (Human-Speech and Music) and the MFCC features extracted from the audio which are used to create an image.

Figure 6: Dataset Representation for the NN model.

Model Accuracy and loss are plotted against Epoch. In initial Epoch Cycles, the model is underfit and it becomes overfit as the number of epochs increases. The ROC Curve for the CNN model is skewed.

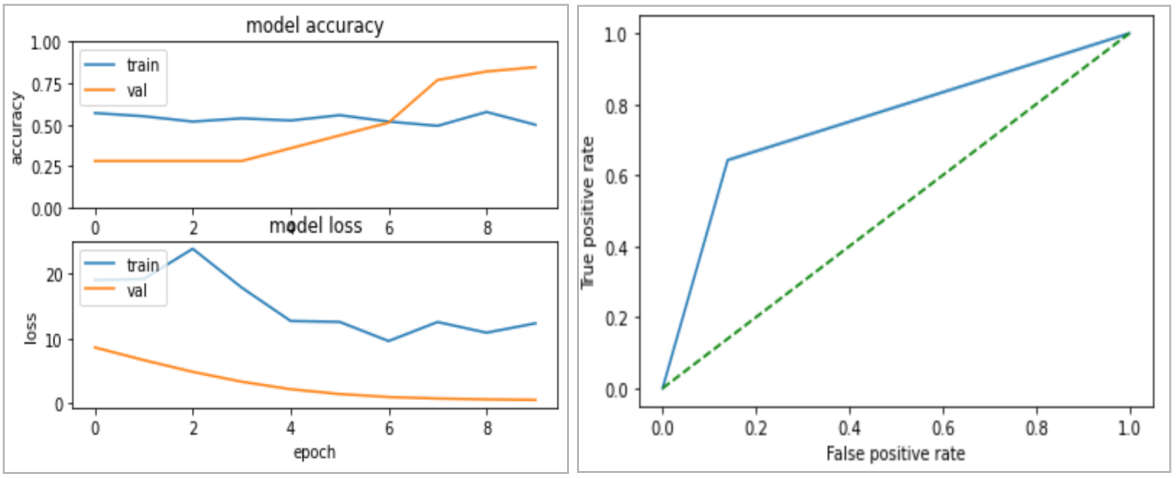


Figure 7: Model Accuracy and Loss plotted against Epoch and ROC Curves

Naive Bayes:

As the number of features increases, the accuracy of the model increases. The model's precision value which describes the true positive cases over the sum of true positives and false positives is also important for our problem statement. As the dataset is small, Gaussian NB performs well, and doesn't require hyperparameter tuning like cross validation, eliminating zero observations, using var_smoothing etc.

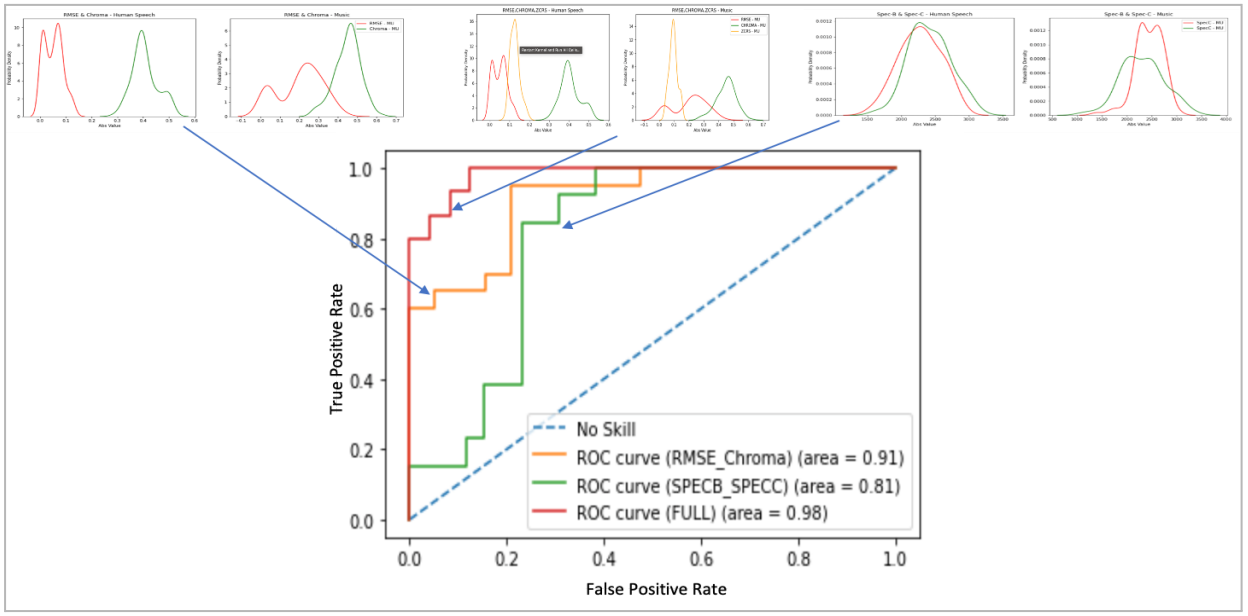


Figure 8: ROC curve for Naive Bayes models with different features

	zcrs	mfccs	chroma	rmse	spec_c	spec_b	rolloff	class_label
0	0.150044	-8.459356	0.388125	0.037607	2612.358748	2139.577794	4620.238729	Human-Speech
1	0.123951	-7.779772	0.405367	0.054338	2159.300344	1991.255180	3995.042509	Human-Speech
2	0.071148	5.206492	0.567012	0.329862	2344.515069	2663.768393	5363.817566	Music
3	0.136798	1.928085	0.449995	0.272583	2985.277462	2790.678112	6261.262139	Music
4	0.126048	-7.293183	0.392444	0.077955	2331.708888	1989.404083	4185.341358	Human-Speech

Figure 9: Mean of different features taken and stored which are actually continuous

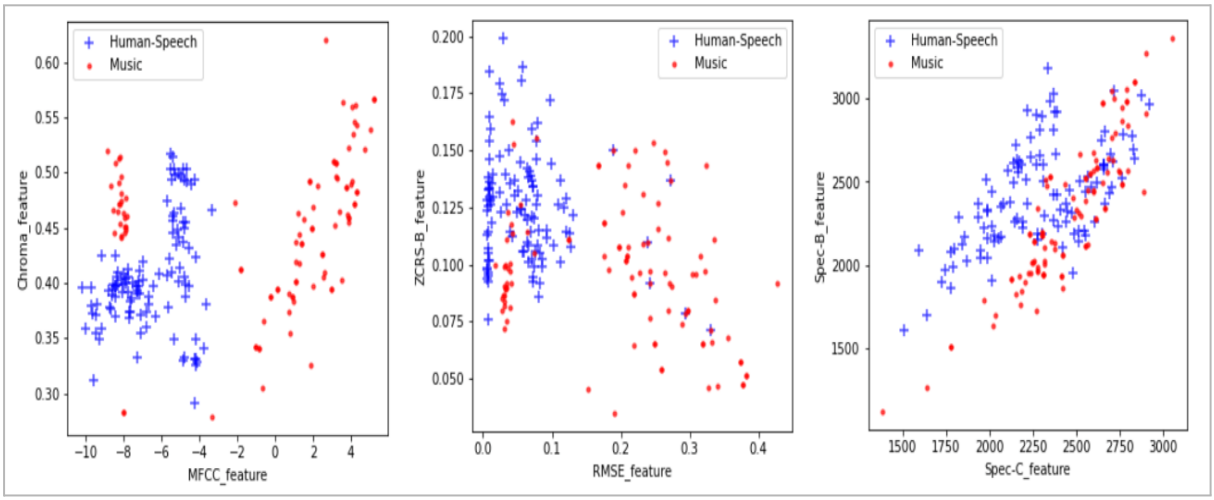


Figure 10: Data Scatter with different features on the x-y axis (mean)

SVM Model:

As SVM model performs well with small dataset therefore the model's performance is really good without using hyperparameters. Radial basis function (RBF) was used as the kernel. The training data accuracy is 97.5% whereas test data accuracy is 95%. Adding more data will make the model overfit.

	zcrs	mfccs	chromagram	rmse	spec_cent	spec_bw	rolloff	class_label
0	0.130065	-8.740104	0.376793	0.050187	2284.931950	1986.540801	4152.608223	speech
1	0.116476	-7.222779	0.395206	0.064112	2149.595660	1881.208683	3934.326210	speech
2	0.141413	-7.290301	0.387690	0.098992	2540.784886	2095.736303	4539.972547	speech
3	0.097891	-7.137390	0.369581	0.074011	1819.815212	1718.981124	3321.563248	speech
4	0.120004	-7.810161	0.405584	0.054381	2076.579681	1892.581596	3839.843448	speech

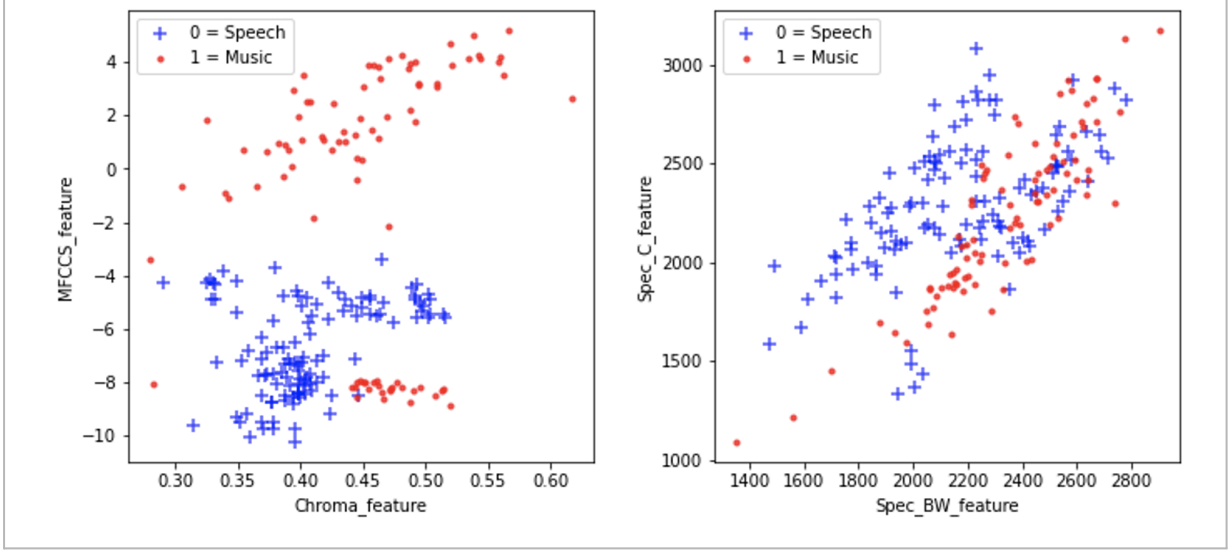


Figure 11: Mean of different features taken and data scatter with different features

ROC curve (receiver operating characteristic curve) shows the performance of a model at all levels. Ideally, we want a classifier with an ROC curve that is close to the top-left corner. Comparing the two ROC curves (figure 12) we can see that the ROC curve of the SVM model is very close to 1 so it is the ideal classifier. AUC (Area under curve) value makes the picture clear. AUC value of 0.99 means that SVM classifier is a very good discriminator.

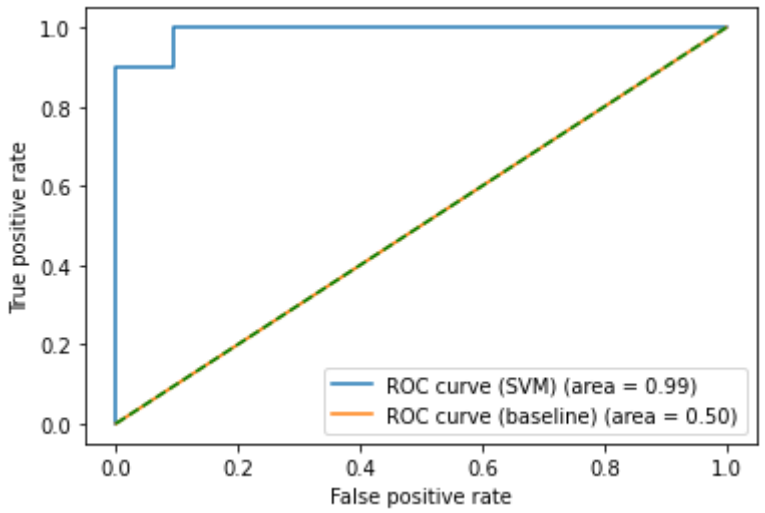


Figure 12: ROC Curve with SVM Classifier

Comparison Parameter	**Baseline Model**	**NN Model**	**SVM**	**Naive Bayes**	**Naive Bayes**	**Naive Bayes**
Features Used	Dummy Classifier (Most Frequent)	MFCC Spectrogram	MFCC, Chroma	RMSE, Chroma	SPECB, SPECC	RMSE, MFCC, Chroma, ZCRS
Accuracy (Model)	0.51	0.84	0.95	0.86	0.69	0.87
True Positives	21	74	21	20	17	22
True Negatives	0	45	18	13	3	12
False Positives	0	25	0	0	10	1
False Negatives	20	12	2	6	9	4
Precision	0.26	0.77	0.96	0.88	0.63	0.88
F1-Score	0.34	0.75	0.95	0.84	0.62	0.86
Recall	0.50	0.75	0.95	0.82	0.62	0.85

Summary

After evaluating the models on the dataset retrieved by our team, we realise that the performance of the classifier depends on the features chosen to represent the audio samples. The best accuracies of each model are : NN 84% , SVM 95% and NB with 87%. As it is evident from the scatter plots, while some features are clearly separable, others tend to make data highly inseparable. Augmenting the data after a certain extent is not required as it could lead to overfitting.

SVM performs better as the MFCC, Chroma feature plots represented in SVM model are independent and separable thereby having clear distinction. A Larger dataset would likely hamper the performance of SVM.

For NN, As represented in Figure 1, the musical and speech audios are quite distinctive even in the waveform pattern therefore the MFCC spectrogram retrieved from the clips are a good fit hence the Neural network also performs well. The accuracy of the training model increases after the 6th Epoch but the model remains under-fitted which can be improved by increasing dense layers in this case.

For NB, since it assumes data to be statistically independent, the only tuning required then is to choose the feature set to distinguish between audio samples more accurately. It is an efficient binary classifier that works well with numerical (and textual) data.

Contributions. This doesn't contribute to the 5 page limit.

Project Proposal:
Written by: (Pallavit, Prishita)
Final Report:
Introduction(Pallavit)
Data Extraction & Processing (Gautam, Pallavit)
Dataset & Features (Prishita, Gautam)

Each Model has been picked by a group member and its corresponding details, results, experiments and conclusions are written by the member.

Group Member	Model
Pallavit Aggarwal	Naive Bayes
Prishita Singh	NN Model
Gautam Thapar	SVM

Summary: (Gautam, Pallavit, Prishita)
Github Link: *https://github.com/singh-prishita/MusicSpeechDiscrimination*