EEC 201 Final Project

Bishane Sagal and Karmvir Thind

March 2024

1 Introduction

The goal of this project is to develop a speaker recognition program. Using 11 given speech samples and 19 speech samples from our colleagues as training data we aim to write a program that will match input test data to the correct student.

2 Approach

Our approach to the problem can be summarized into three steps:

1. Perform feature extraction on the audio data
2. Develop/Choose an algorithm to learn from the features and fit a model
3. Classify new audio data using the developed model

2.1 Feature Extraction

For feature extraction we made use of the Mel Frequency Cepstrum Coefficients (MFCCs). We followed a the mechanical procedure for producing the MFCCs: Compute the Short Time Fourier Transform of the input data (STFT), convert the magnitude of the STFT (to eliminate complex numbers) to mel-scale using appropriately spaced filter banks, take the log of the mel-scale STFTs, and perform the Discrete Cosine Transform (DCT) on these values to produce the MFCCs.

2.2 Algorithm Selection

We tested various models available through Python's ScikitLearn library and also developed our own implementation of the recommended Linde-Buzo-Gray (LBG) algorithm. Ultimately we landed on the [INSERT FINAL MODEL HERE]. We tested various models using for loops to sweep through various parameters and hyperparameters. Using the classification output we tuned the model to achieve our results of [insert accuracy here]

2.3 Model Prediction

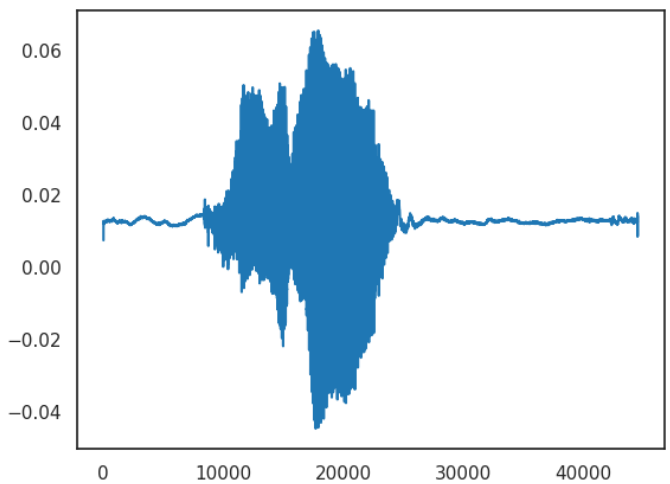
Assessment of our model was done using features of the pre-built models from Python's Scikitlearn or in the case of our LBG algorithm, we displayed a confusion matrix to make our results easy to determine.

3 Tests

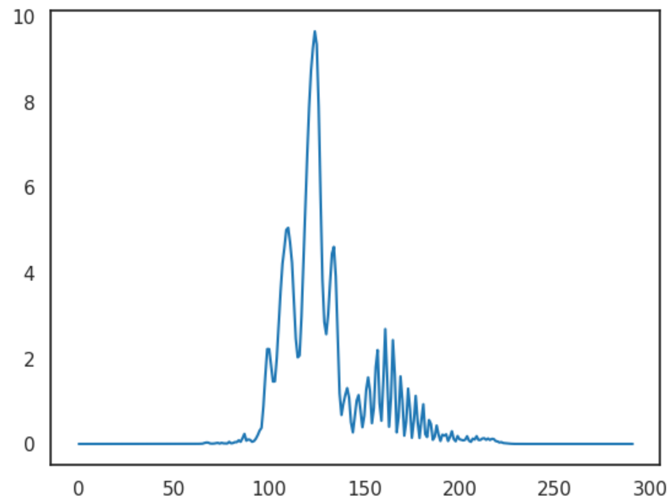
3.1 Test 1

Auditory Performance rate was 7 out of 11 students guessed correctly.

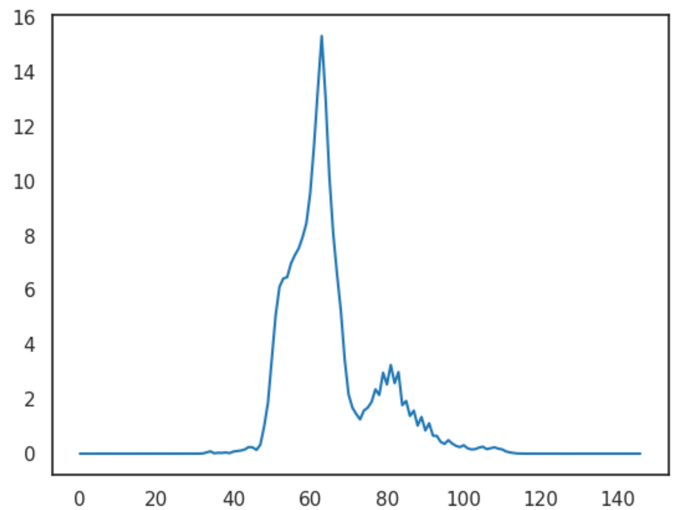
3.2 Test 2



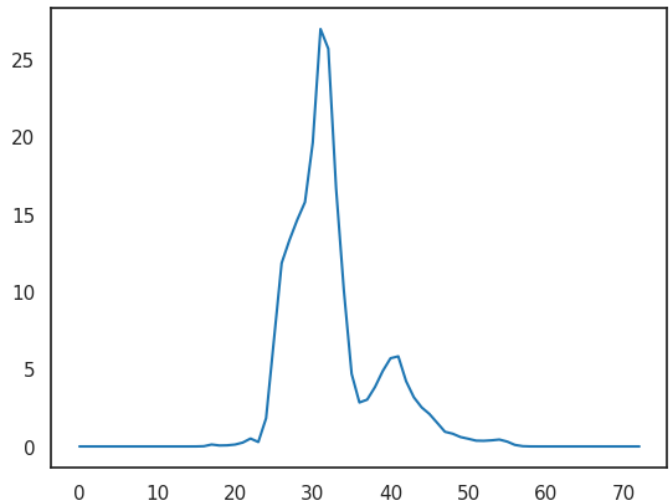
The sampling rate of our load function is 22050, giving us 256 samples in 11 milliseconds.



Periodogram with N = 128

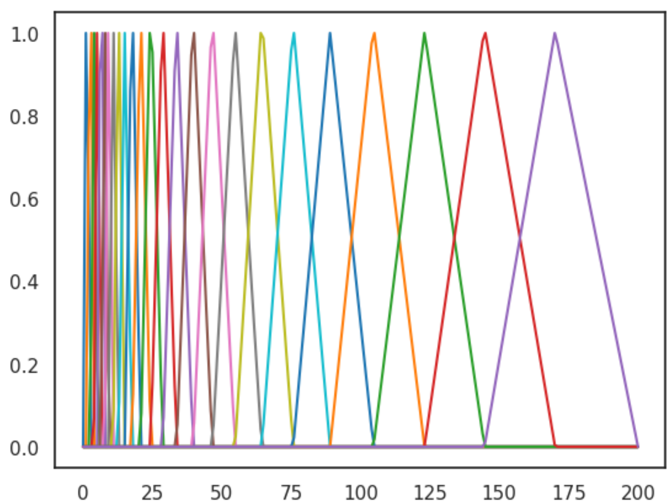


Periodogram with N = 256

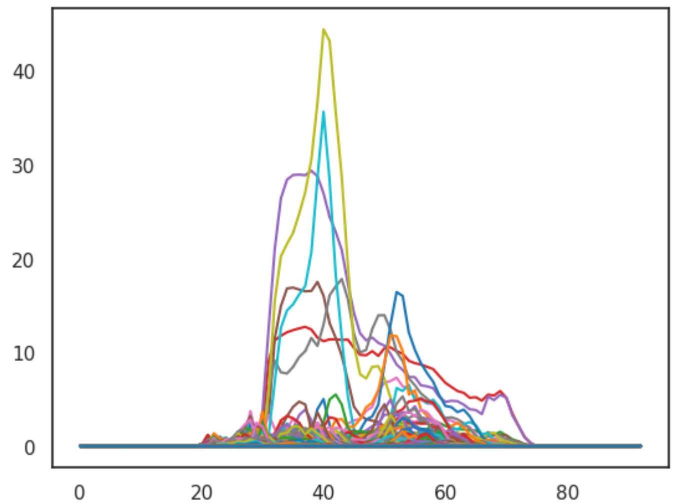


Periodogram with N = 512

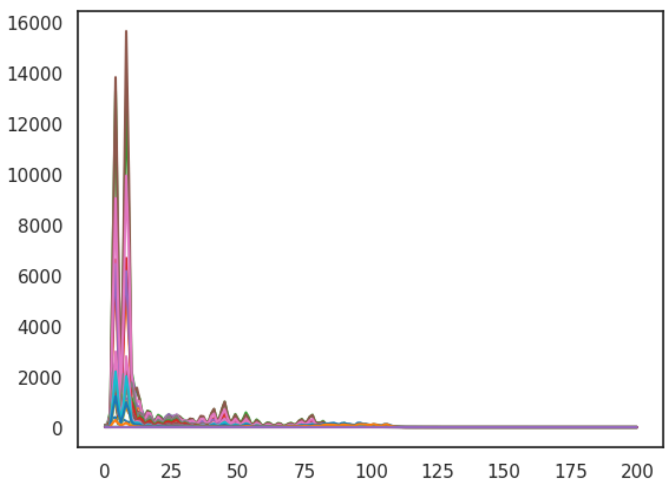
3.3 Test 3



Ideal Mel Spaced Filter Banks



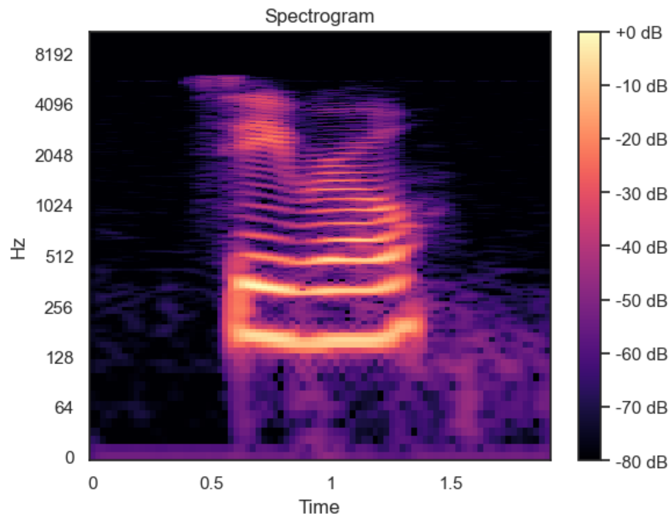
Spectrum of an audiofile with no mel frequency wrapping



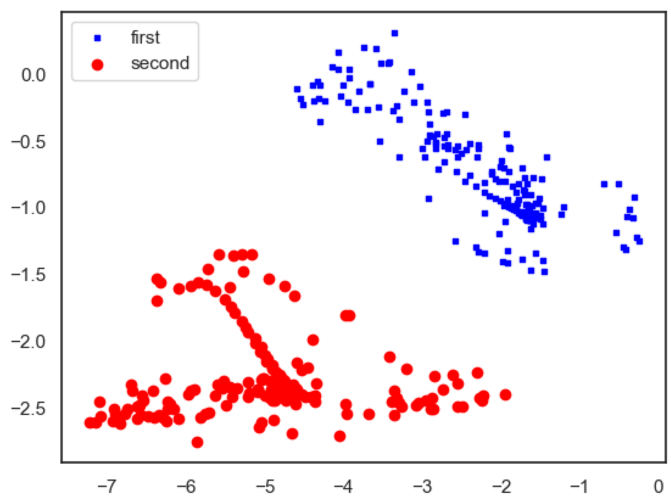
Spectrum of audio file with mel frequency wrapping

3.4 Identifying Points of High Energy

As seen below, we have high intensity points in the range of 0.5 to 1 seconds. The highest energy points are with the 128-512 frequency band. The intensity plotting is in accordance with our raw audio file coefficients because our largest amplitudes lie around the 20000 sample which corresponds to approximately 1 second.



3.5 Test 5



Training Data MFCC vectors plotted in a 2-D plane