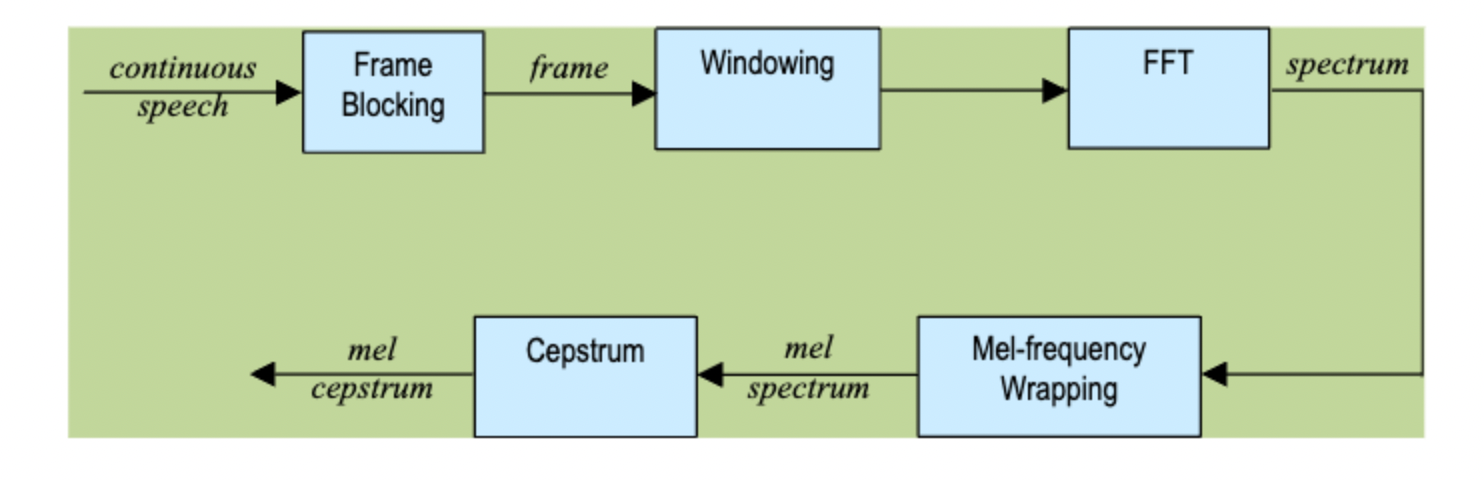
Abstract

In our project, we explored a technique for recognizing individuals based on their voices, employing a specific method known as the LBG algorithm. Our aim was to identify and map unique voice features, utilizing Mel-frequency cepstral coefficients (MFCCs) as the primary tool for this task.

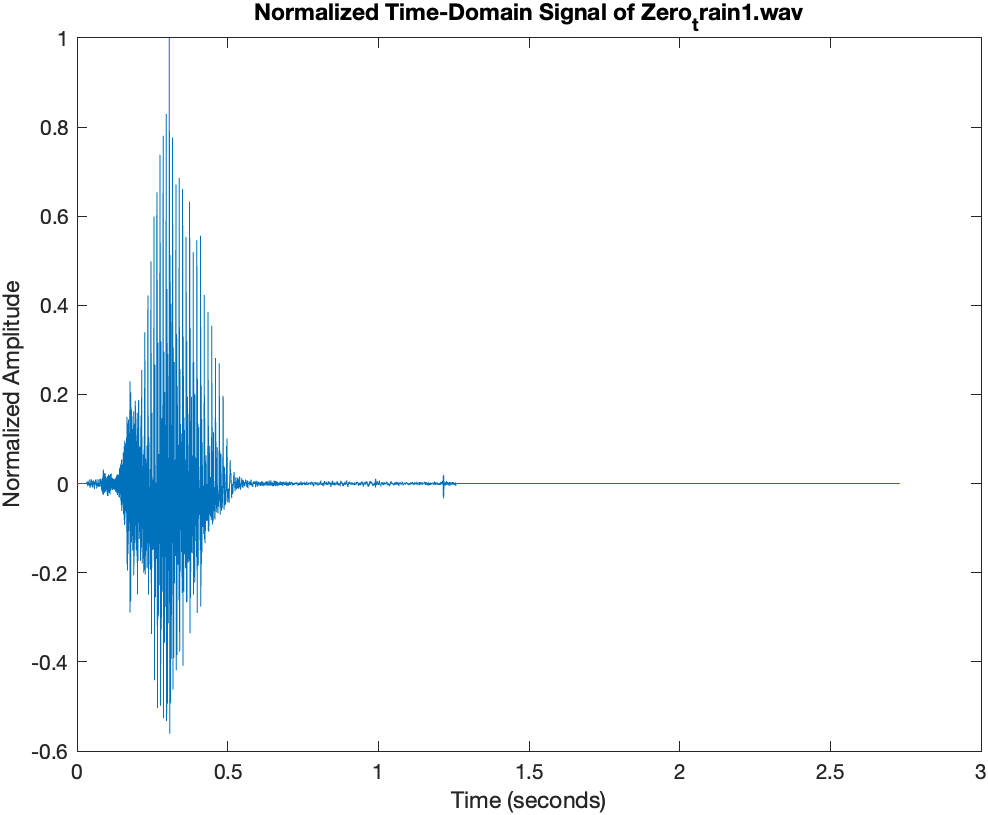
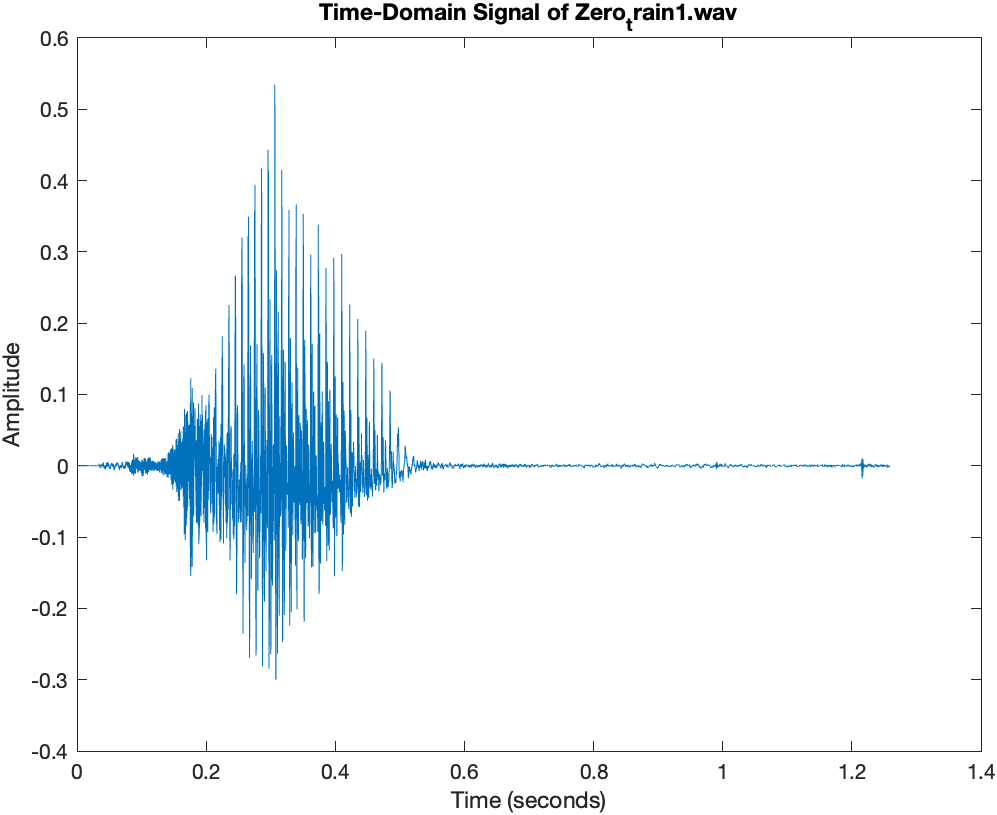


*Figure – MFCC*

The process began with the initial preparation of voice recordings, which involved cleaning and structuring the data through segmentation and the application of the short-time Fourier transform (STFT). Following this, the LBG algorithm was implemented to categorize the voice data into distinct groups, according to their shared characteristics. This categorization facilitated the creation of a distinctive codebook for each participant, effectively capturing the essence of each individual's voice.

To evaluate the efficacy of our system, we conducted tests using 18 voice samples collected during class sessions. Impressively, our system achieved successful identification for 13 of these samples, achieving a recognition rate of 72.2%. This outcome notably surpasses the average human recognition rate, which typically stands at about 50%, demonstrating the potential of our approach in the field of voice recognition.

We may need some backgrounds like the normalization part...



Tests

test1:

benchmark: human can recognize 50% of audios

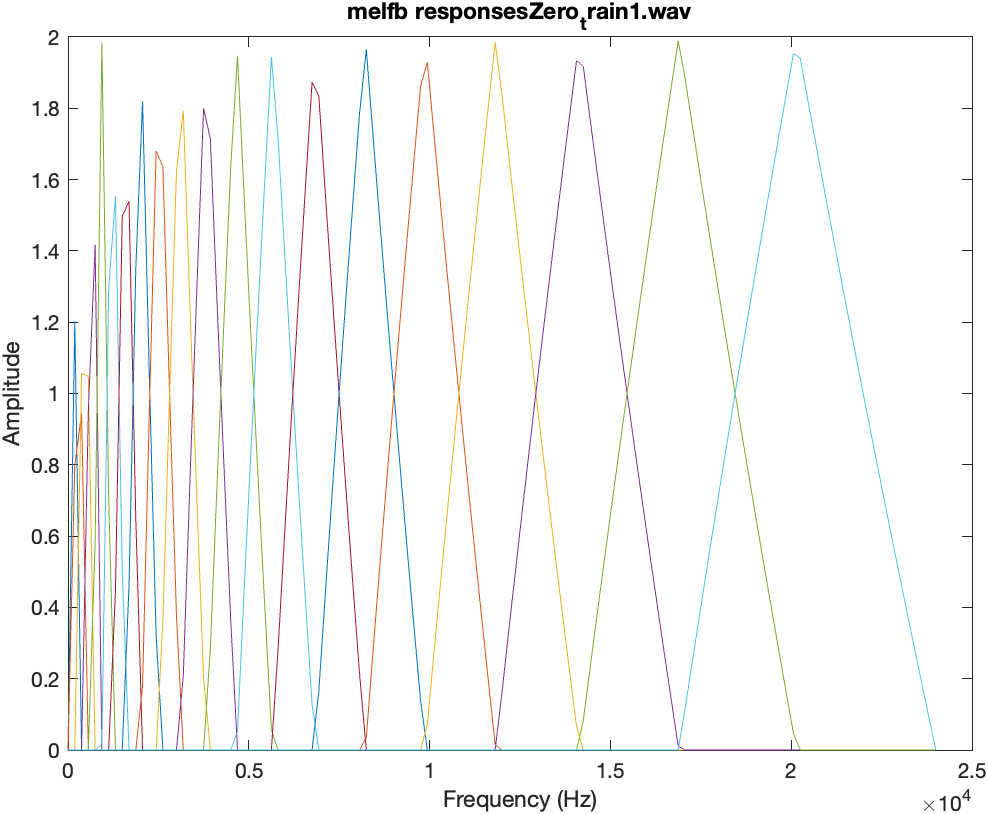
test2:

sampling rate: 48000Hz

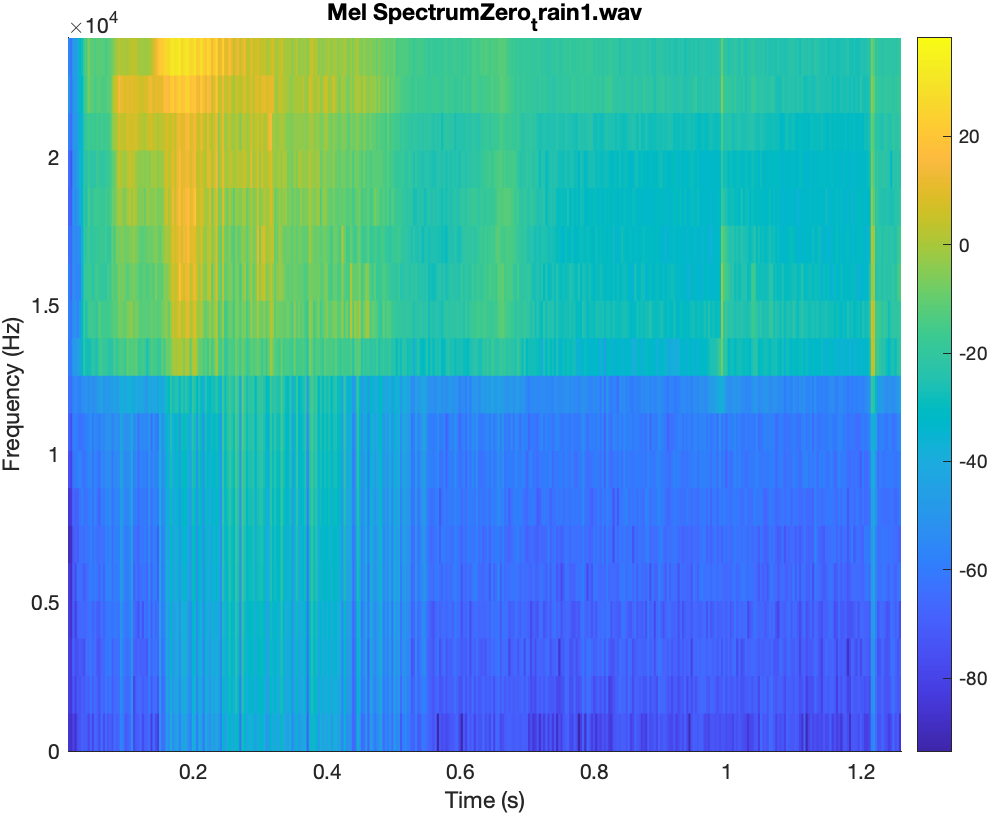
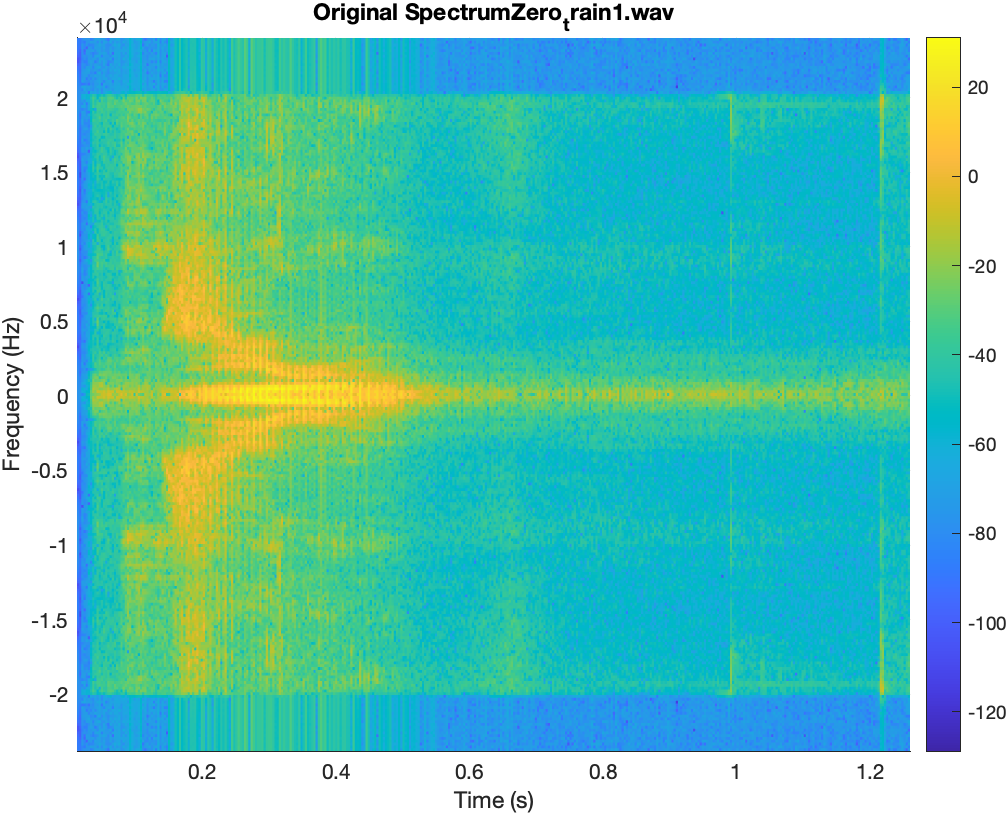
millisecond numbers contains in a block of 256 samples: 30 msec

we choose a 256 samples frame and 100 samples overlap based on the length of each audios. Larger frame increases the recognize speed but reduce the accuracy.

test3:



*Figure – melfb responses*

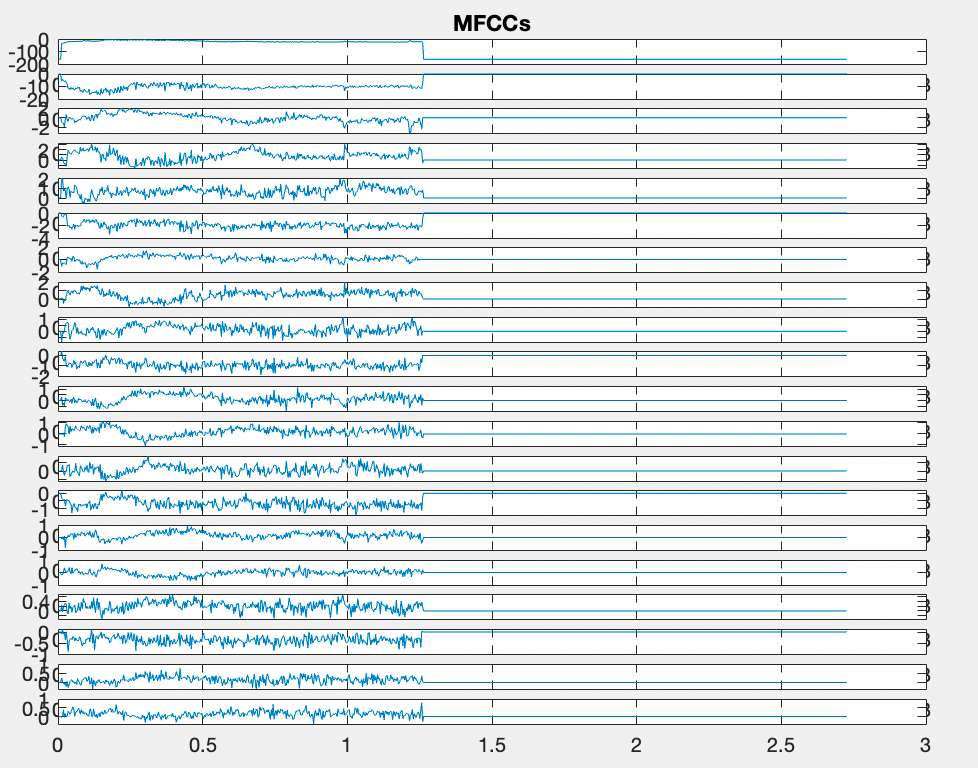


*Figure – original spectrum Figure – spectrum after melfb*

melfb.m : Since psychophysical studies discovered that human perception of sound frequency does not follow a linear scale, we use melfb function to scale frequencies. The melfb function takes the number of desired Mel filters (K), the FFT length (n), and the sampling rate (fs) as inputs and generates a matrix containing K filters to represent the Mel filter's responses. We can apply these filters to FFT results to generate spectrums that emphasize frequencies in a way that aligns with how people perceive sound.

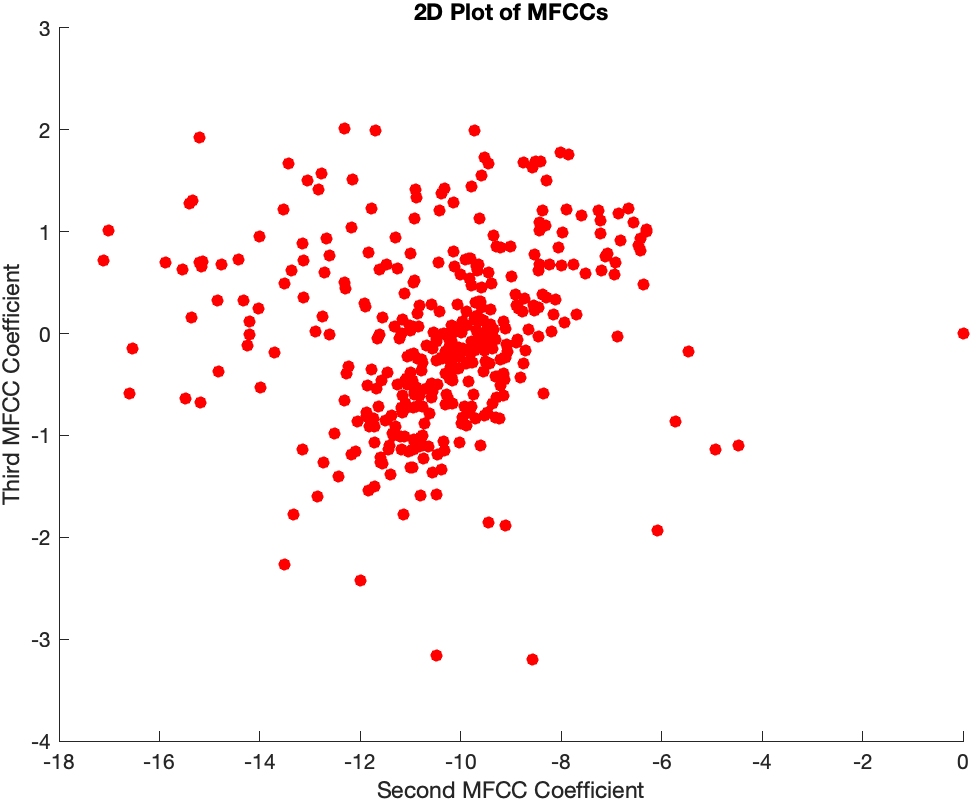
test4:

We bulit a function called mfcc\_test, it works well.



*Figure – MFCCs*

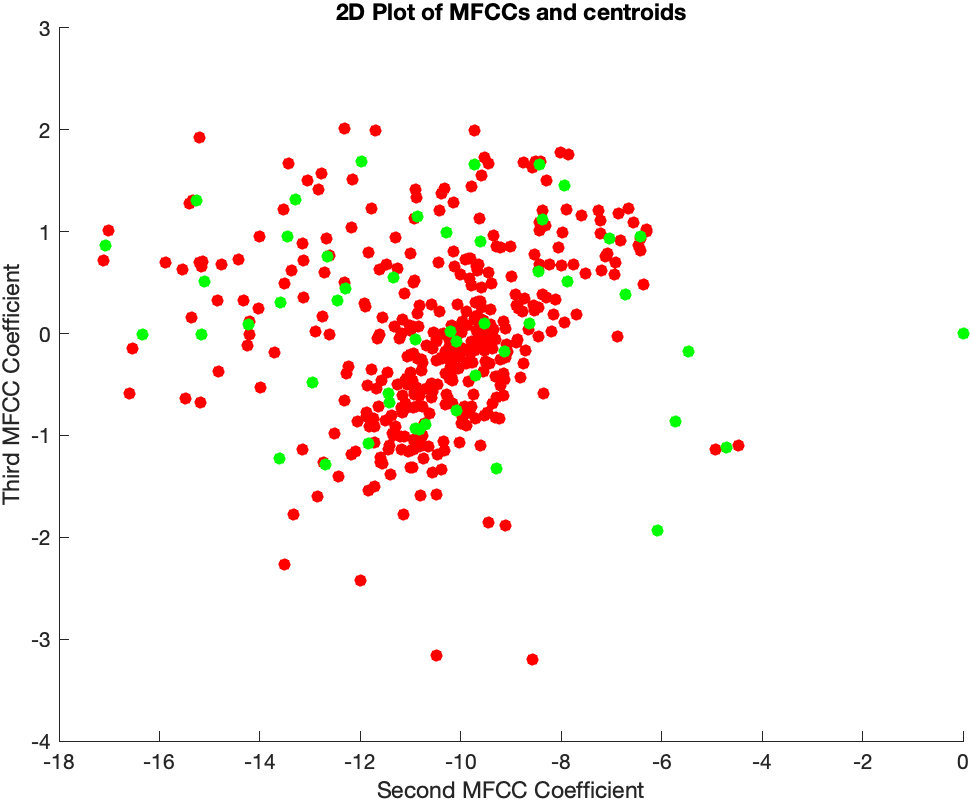
test5:



*Figure – 2D plot of MFCCs*

They are in clusters but not as clear as the figure shown in the class.

test6:



*Figure – 2D plot of MFCCs and centroids*

This is the codebook of the first train audio. Red points represent original MFCCs, green points represent generated centroids.

test7:  
Our program successfully identified 13 out of 18 audio samples, achieving a recognition rate of 72.2%. This performance exceeds our benchmark of 50%. To enhance the recognition accuracy, we could refine the LBG algorithm by incorporating k-means to determine centroids more effectively. Additionally, the consistency of the training audio samples, particularly their duration, has a notable impact on the results. Ensuring similar lengths in the time domain for all training samples could further improve accuracy.

We expanded our testing by recording two additional audio samples, which were correctly recognized by our system, indicating that our system is generalizable.

test8:

With a sampling frequency of 48000Hz, we tested the robustness of our system by applying a notch filter with a stopband from 10000 to 10100Hz to our training audios. The recognition rate remained at 72.2%, demonstrating good robustness against this type of frequency interference. However, expanding the stopband width to 1000Hz will affect accuracy, indicating our system’s robustness limitation.

test9:  
We expanded our voice sample set by recording two friends saying the word "zero" and tested our system's accuracy with these new inputs. Both voices were successfully recognized, improving our overall accuracy to 83.3%, with 15 out of 18 samples correctly identified.

test10:

For all zeros: 72.2%, 13/18

For all twelves: 55.6%, 10/18

test11:

We also incorporated yes.wav, no.wav, yes we are, and no you are not from computer assignment 6 to further evaluate our system. The result shows that our system can recognize individuals regardless of the words spoken. However, this finding requires additional validation considering the limited range of samples tested.

Group division:

Yuyang Jin: I mostly handle tests except for test 6 LBG algorithm, including preprocesses before and after LBG, figure plots, and data samples tests. I also gave some code attempts, including k-means and the shortest distance algorithm, and helped Joey debug his code.

Joey: report your efforts...