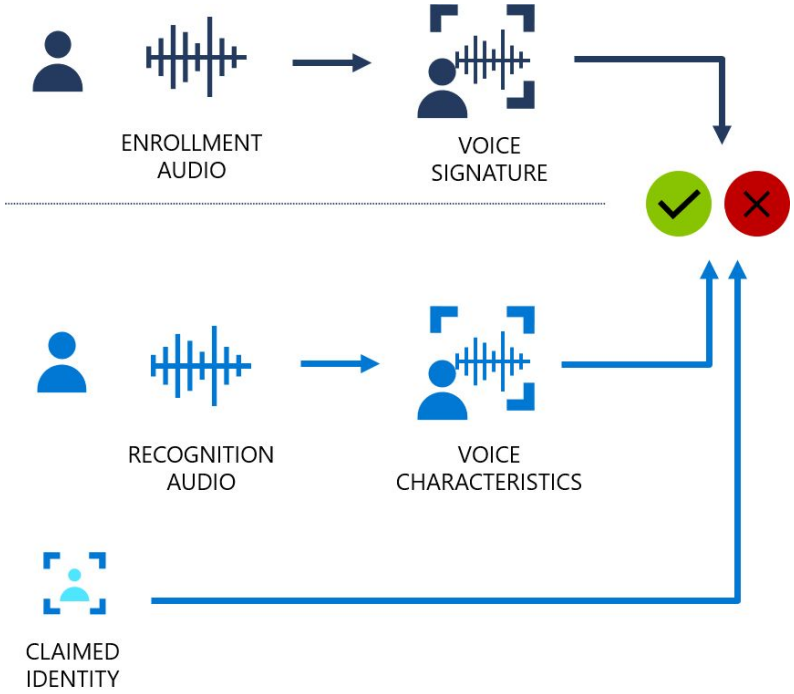
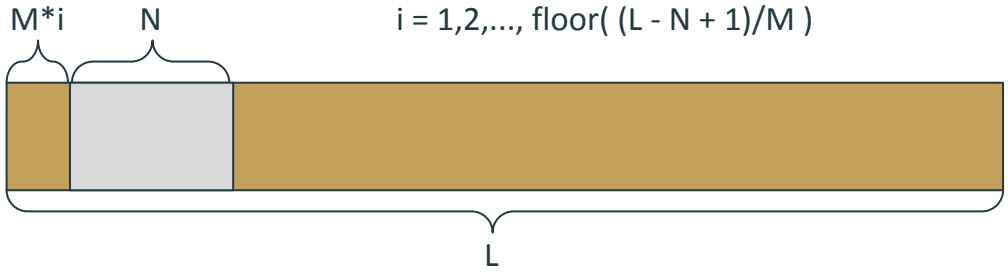
Speaker Recognition

Randall Fowler and Conor King
Winter 2024

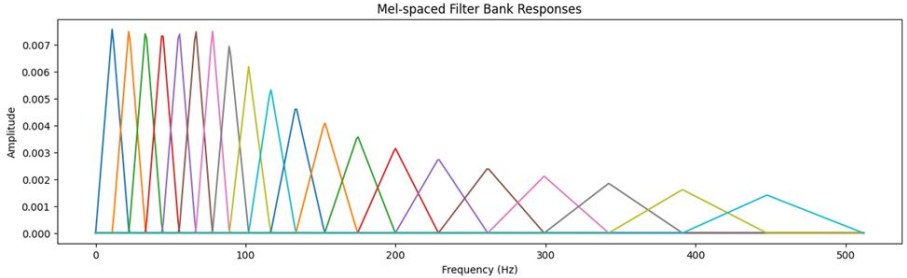
Speaker Recognition



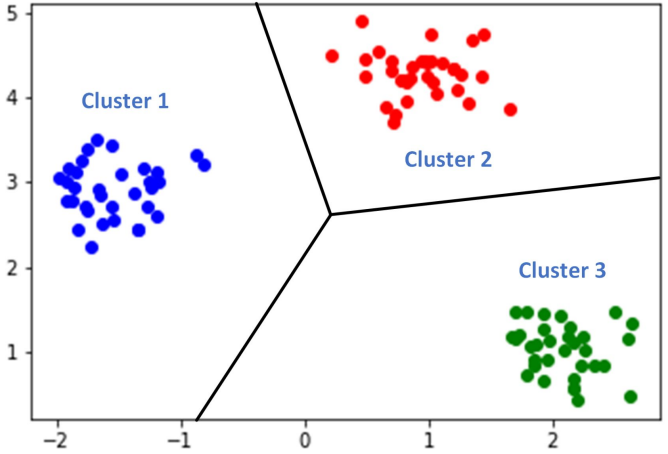
Methodology



- Feature Extraction - Calculating Mel Frequency Cepstrum Coefficients (MFCC)
- Each frame of audio signal will require windowing to remove ringing.
  - Hamming, Hanning, Blackman, and Bartlett
- Mel Frequency Spectrum
  - Apply Frequency Transform and triangular weights to get a Mel-Spectrum.
  - Apply DCT on the Mel-Spectrum to get Cepstrum Coefficients.
    - Number of Coefficients (K) is equal to the number of triangular weights used to get the Mel-Spectrum.
- Resulting data related to an audio file will have length of number of frames and K dimensional.



Codebooks



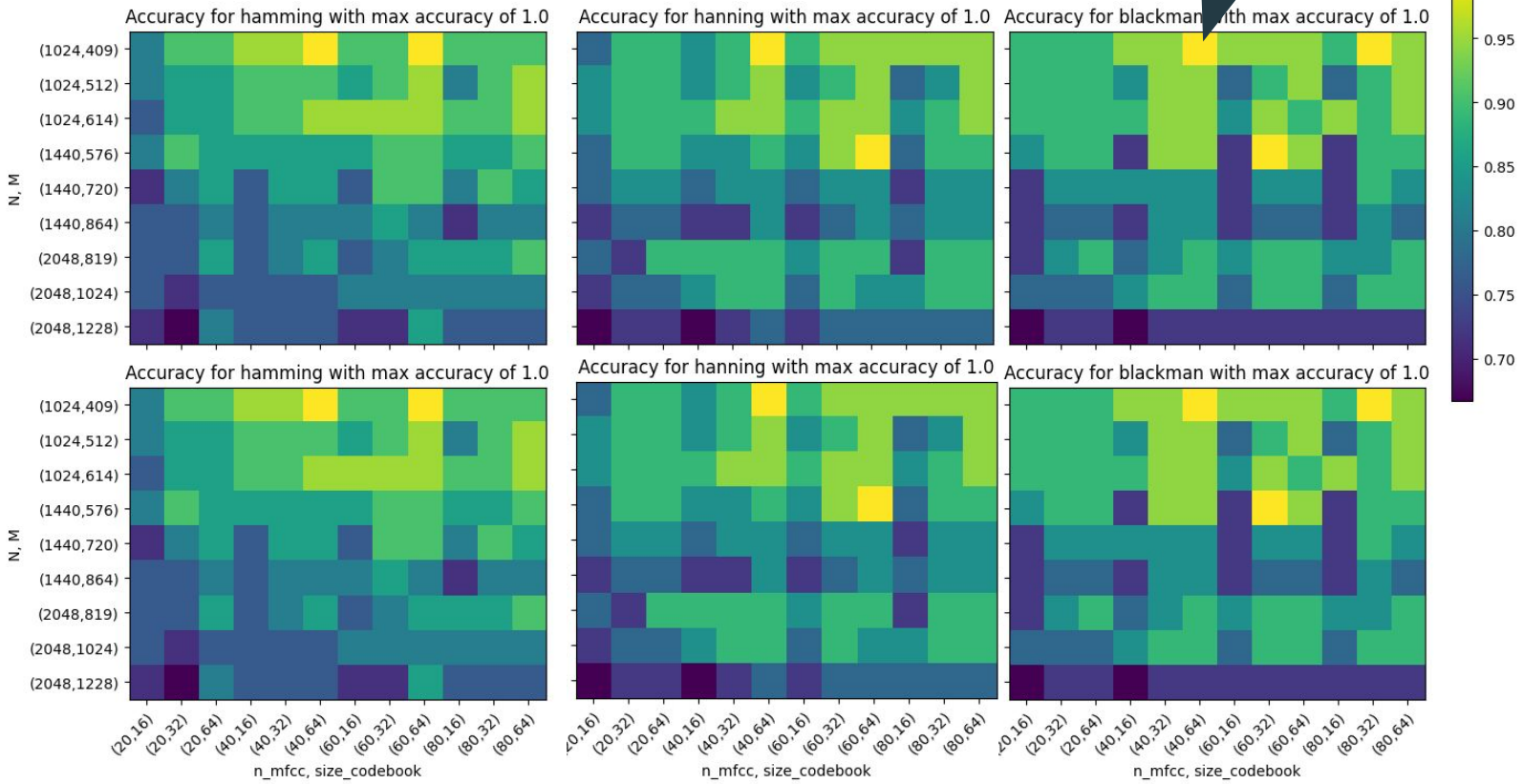
- To classify speakers in audio files, centroids will be calculated on training data.
- In the data space containing the MFCC, clusters are formed by setting an initial centroid, moving it to the mean, and splitting.
- The number of centroids will be a hyperparameter, but the clustering algorithm will move centroids to the mean of the closest data points.
- Testing audio files can be classified by getting the MFCC data and comparing the distance to the centroids of different codebooks.
    - Prediction will be the codebook with the smallest distance.

Hyperparameter Sweep

Only optimal point in Bartlett

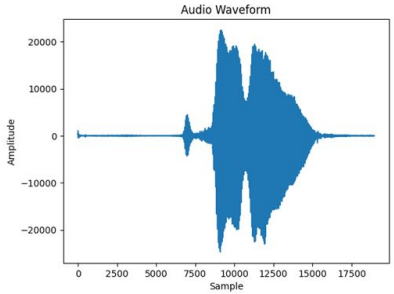
TWELVE:

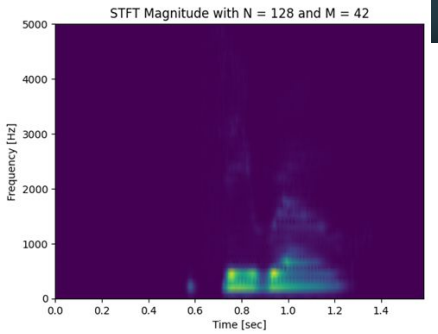
ZERO:

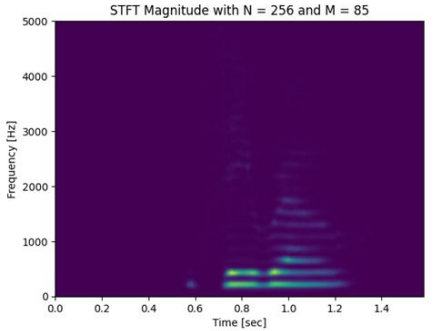


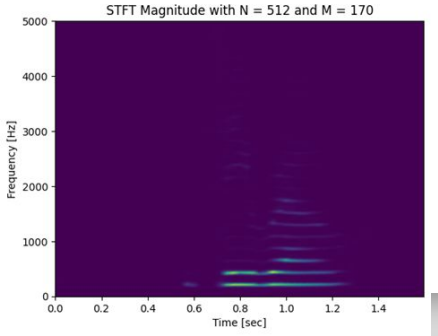
Tests 1 and 2

- Test 1
  - Randall tested, Conor guessed
    - 3/8 correct => 0.375
- Test 2
  - Sampling rate: 12kHz
    - 256 Samples => 21.3 ms.

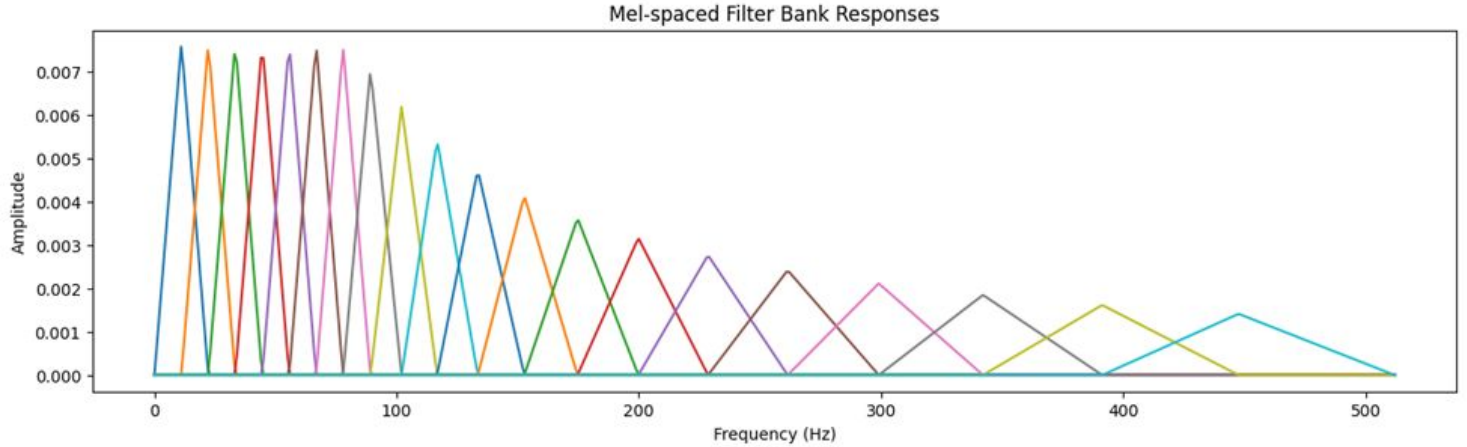


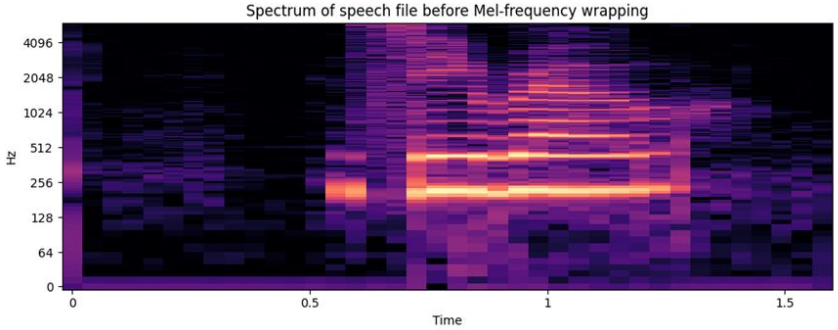


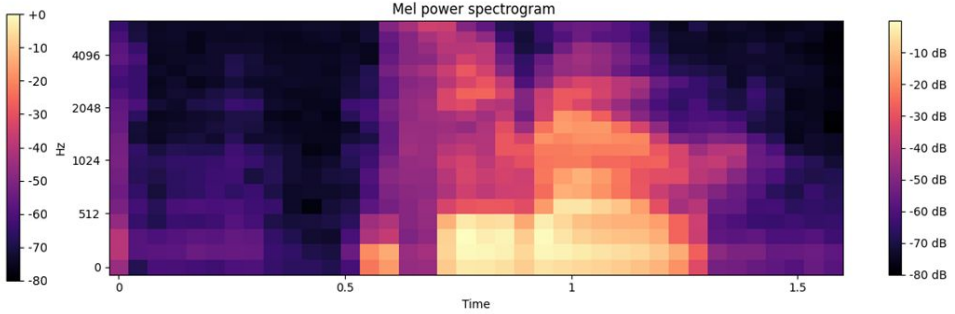




Tests 3 and 4

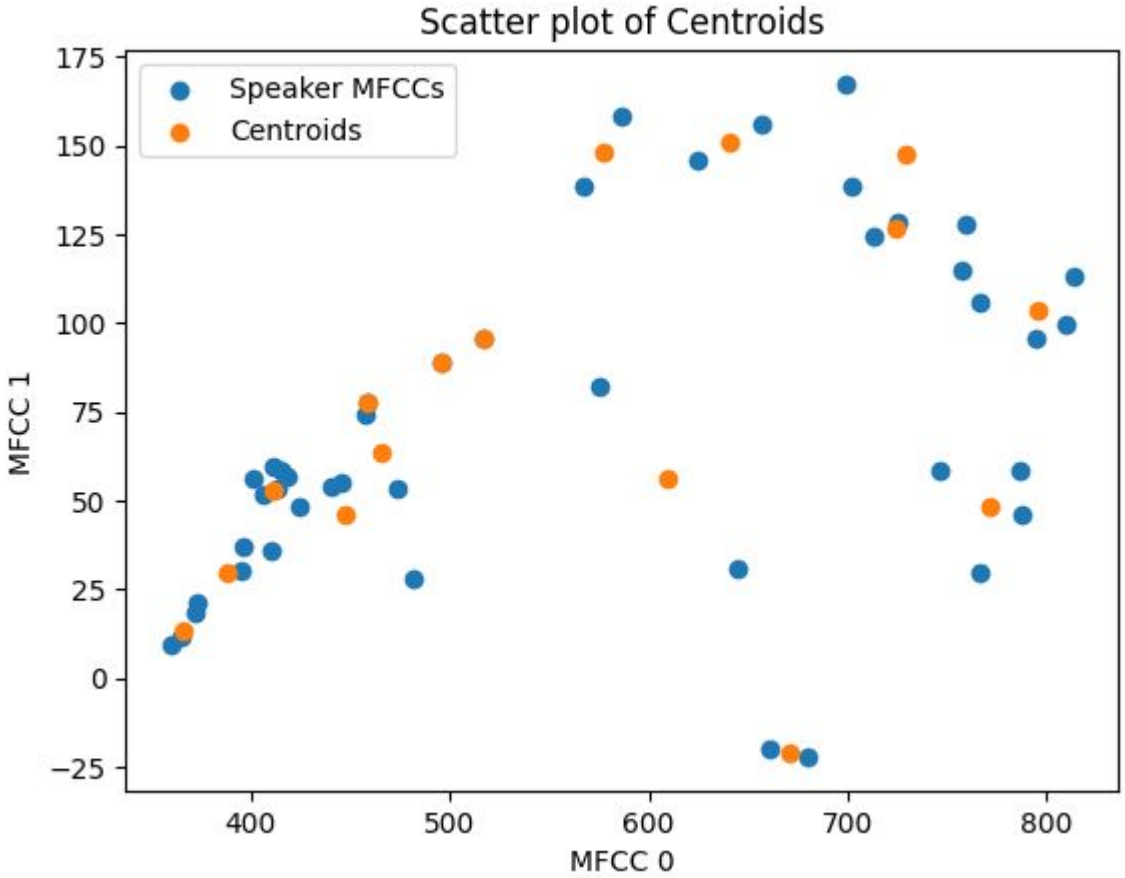

Mel-spaced Filter Bank Responses
Amplitude
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
100
200
300
400
500
Frequency (Hz)


Spectrum of speech file before Mel-frequency wrapping
Hz
4096
2048
1024
512
256
128
64
0
0
0.5
1
1.5
Time


Mel power spectrogram
Hz
4096
2048
1024
512
0
0
0.5
1
1.5
Time
+0
-10
-20
-30
-40
-50
-60
-70
-80
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Tests 5 and 6

Finding clusters in the FMCCs, and creating centroids.



Tests 7 and 8

Hyperparameters:

N = 256

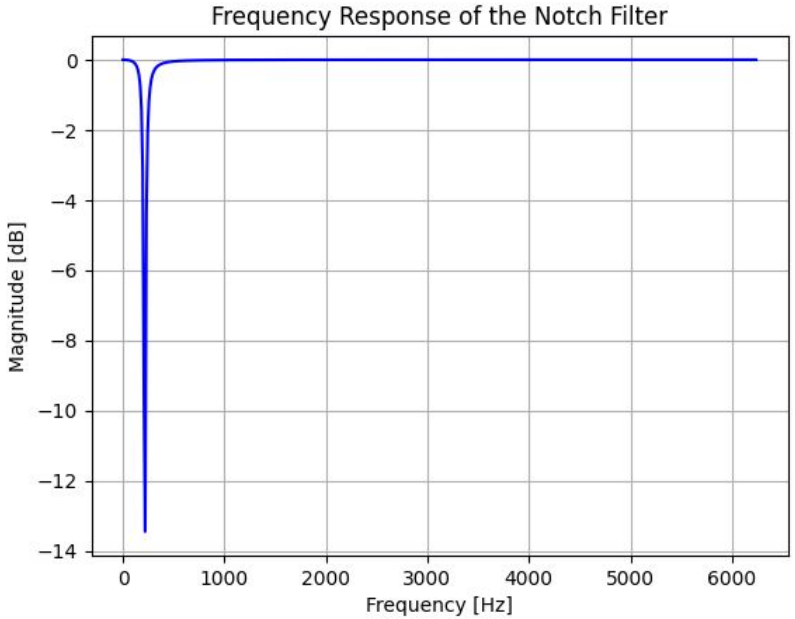
M = 100

n_mfcc = 40

size_codebook = 64

window = Hamming

Accuracy on the given datasets: 8/8 correct = 1.0



Notch at 215: 1.0
440: 1.0
1000: 0.75
6000: 0.875

Tests 9 and 10

- Test 9: Given and 10 random 0s
  - Accuracy: 0.83
- Test 10
  - Question 1: 12s
    - Accuracy: 1.0
  - Question 2:
    - a) All files
      - Accuracy: 0.886
    - b) Determining “zero” or “twelve”
      - Accuracy: 1.0

Conclusion

Our system is effective at distinguishing voices.

Robust to notch filtering.

Non-homogeneous data (with different sampling rates) causes performance to suffer somewhat.

Improvements: smoothing ambient noise, trimming data ends, normalizing energy