Speaker Verification

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

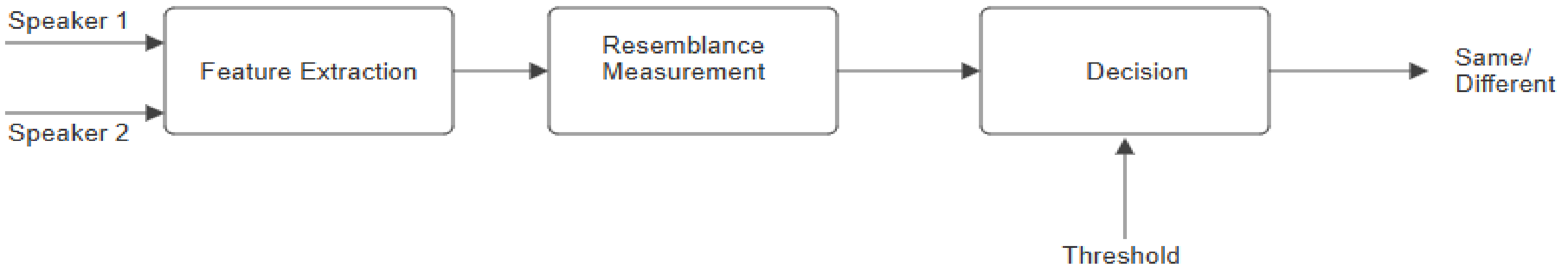
Aniket Dwivedi | MT24208
Bikrant Bikram Pratap Maurya | MT24116
Nakul Panwar | MT24057
Vardhana Sharma | PHD242001

Problem Statement

- Speaker verification enables secure and efficient voice-based identity authentication, which is critical for access control and voice-activated systems. Our project uses machine learning techniques to verify speakers by analyzing audio samples from different speakers to determine whether they belong to the same individual or other individuals.

Problem Statement

- The challenge lies in applying Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) techniques to identify speakers in the dataset accurately. This contributes to the progress of speaker recognition technology, with applications in diverse fields, including security and accessibility for individuals with physical challenges.



About the Dataset

1. There are 25 speakers, and the total count of audio files is 2944, which is a total of 24330.307029478376 seconds.

STATISTICS	In seconds
mean	8.264371
std	5.873632
minimum	3.960091
25%	4.880091
50%	6.420091
75%	9.250091
maximum	69.040091

	Speaker	Total Duration (seconds)	Min Duration (seconds)	Max Duration (seconds)	Number of Files
0	id10278	1228.531688	3.960062	29.760063	187
1	id10284	745.325625	4.000063	30.960062	90
2	id10289	757.925437	4.040063	44.600062	87
3	id10294	853.928625	3.960062	17.400063	138
4	id10281	805.165250	4.000063	38.600062	84
5	id10287	347.723000	3.960062	18.920063	48
6	id10277	418.084187	3.960062	12.600062	67
7	id10290	1094.608562	3.960062	22.400063	137
8	id10275	563.844625	3.960062	21.440062	74
9	id10272	337.083125	4.000063	18.040063	50
10	id10280	679.084188	3.960062	32.720062	67
11	id10276	1674.371562	3.960062	31.880063	185
12	id10285	735.525813	3.960062	29.280062	93
13	id10283	3218.814562	3.960062	69.040063	233
14	id10286	1395.009312	3.960062	35.360062	149
15	id10273	1902.055000	3.960062	42.760062	240
16	id10288	447.163000	4.120063	45.120063	48
17	id10282	683.485250	3.960062	33.560063	84
18	id10271	438.844563	4.000063	13.880062	73
19	id10274	316.883375	3.960062	12.880062	54
20	id10270	1044.849875	3.960062	18.800062	158
21	id10279	453.803937	3.960062	25.280062	63
22	id10292	1710.496562	3.960062	19.520063	265
23	id10293	1626.412125	3.960062	32.400062	194
24	id10291	851.204750	4.000063	31.280062	76

Exploratory Data Analysis

1. Data preprocessing :

Preprocessing: Audio Normalization: Ensuring all audio files are sampled at a consistent rate (e.g., 16000 Hz) to maintain uniformity.

Segmentation: Audio files are split into smaller, fixed-length segments (3 seconds and 8 sec). This allows for handling variable-length audio files while ensuring sufficient data for training.

Padding: Shorter audio segments are padded with zeros to meet the required length. This avoids issues with variable-length input and ensures consistent feature extraction across all segments.

Noise Reduction: Unwanted noise is minimized using techniques like band-pass filtering, which improves the quality of the extracted features.

Feature Extraction: Key features like MFCCs (Mel-frequency cepstral coefficients) are extracted from each audio segment to represent the speaker's voice characteristics.

Methodology and Feature Extraction

2. Feature Extraction:

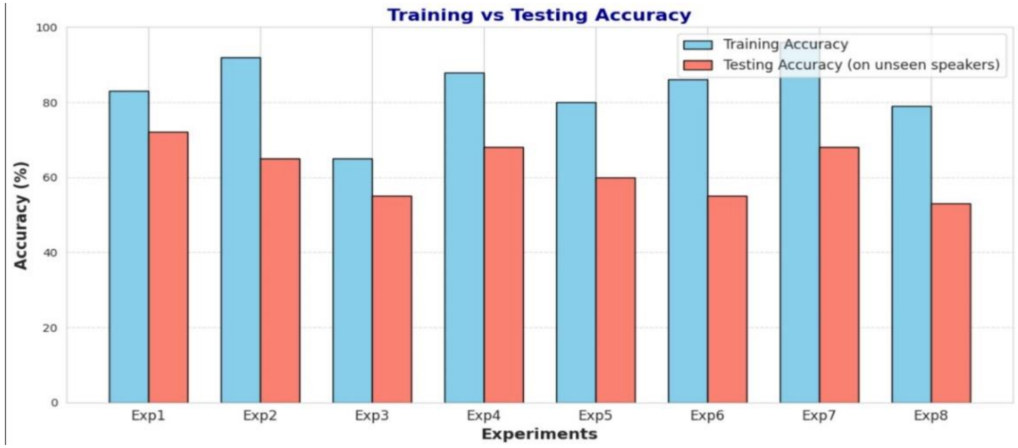
- In this speaker verification project, we utilized a comprehensive set of audio features, including MFCC, chroma, spectral contrast, and spectral centroid, along with their statistical aggregates like mean values (e.g., MFCC_mean, spectral_centroid_mean). Additionally, advanced features such as pMFCC, mel spectrogram, and pitch-related features like pitches and pitch_mean were extracted to capture the unique characteristics of each speaker's voice. These features enabled robust representation of audio data for accurate classification, facilitating the verification of whether two audio samples belong to the same or different speakers.

3. Model Training

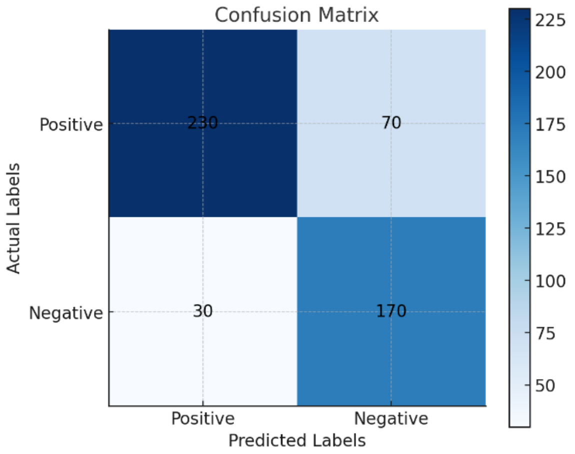
- Gaussian Mixture Models (GMMs) for each speaker using their audio data. It processes audio files, extracts MFCC features, trains a GMM for each unique user Ids, and stores all models in a dictionary. The combined models are saved as a serialized .pkl file, with progress logged throughout the process.

Results

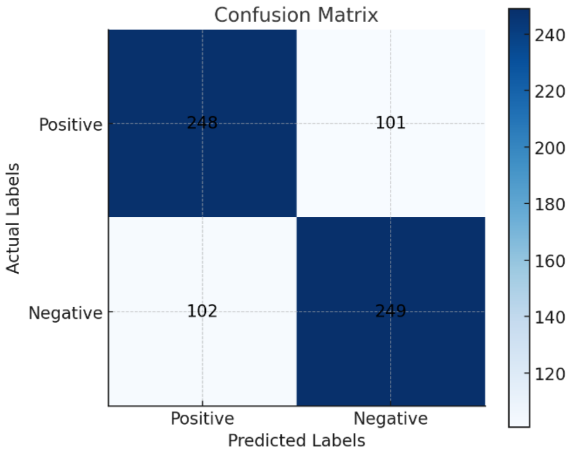
- Among the evaluated configurations, on training data we are getting accuracy of **91%.**
- By using GMM model taking parameter (n component = 64).



- Among the evaluated configurations, on unseen data we are getting accuracy of **71.7%.**
- By using GMM model taking parameter (n component = 64).



Seen Data



Unseen Data

Conclusion

- The traditional machine learning models trained on the dataset yielded low accuracy for speaker verification, highlighting the limitations of such approaches for this task. Factors like background noise, pitch variations, and dataset inconsistencies may have contributed to the suboptimal performance. This underscores the need for advanced techniques, such as deep learning models or feature engineering, to capture the complexities of speaker characteristics better and improve verification accuracy in future implementations..

SPEECH VERIFICATION