

Speaker Verification



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

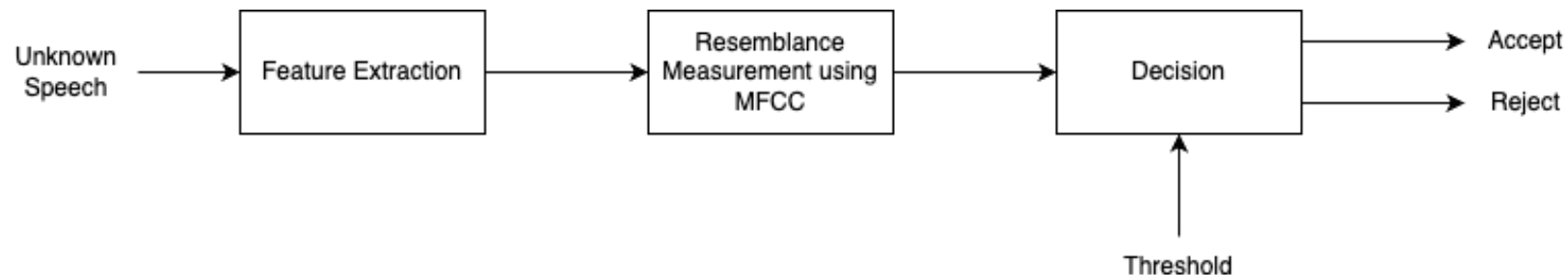
BY
Mayank Gupta



Problem Statement



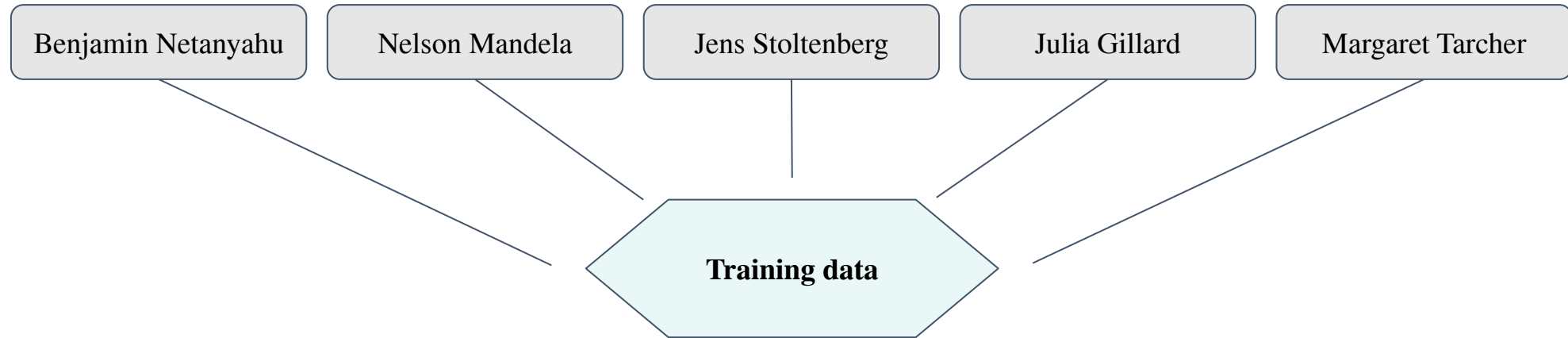
- Speech recognition facilitates identity verification and access control, enabling voice-activated services. Our task focuses on Speaker Recognition, analyzing speeches by five leaders: Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Thatcher, and Nelson Mandela.
- Applying Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) techniques to accurately identify speakers in the dataset. 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 **5 speakers**, and each of their speeches has been divided into 1500 audio chunks (0.wav to 1500.wav), each lasting for **1 second** at a sample rate of 16000.



2. The **background_noise folder** consists of audio recordings that aren't speeches but capture various ambient sounds present in and around the speaker's environment, like laughter or applause from the audience. Within the **other** folder, there are additional types of noise present, such as pink noise, among others.



Exploratory Data Analysis



1. Data preprocessing :

At first, background noise was segmented into one-second chunks and integrated with speech audio using additive mixing. There is background noise in 2 folders, having audio files of different lengths. We have made one-second noise clips from the original audio file, with the same sampling rate as that of the speech files.

2. Use of ZCR over RMS for silence removal :

Utilizing Zero Crossing Rate (ZCR) thresholds instead of Root Mean Square (RMS) is crucial for accurate silence elimination in noise preprocessing.

<u>Speaker</u>	<u>ZCR variance Threshold</u>
Benjamin Netanyahu	425
Jens Stoltenberg	500
Julia Gillard	700
Margaret Tatcher	300
Nelson Mandela	650

Methodology and Feature Extraction



3. Feature Extraction:

- Feature extraction involved MFCC, delta, and delta2 from audio files
- Normalization using dBFS is essential for maintaining consistent loudness across diverse audio samples
- Variations in loudness are standardized

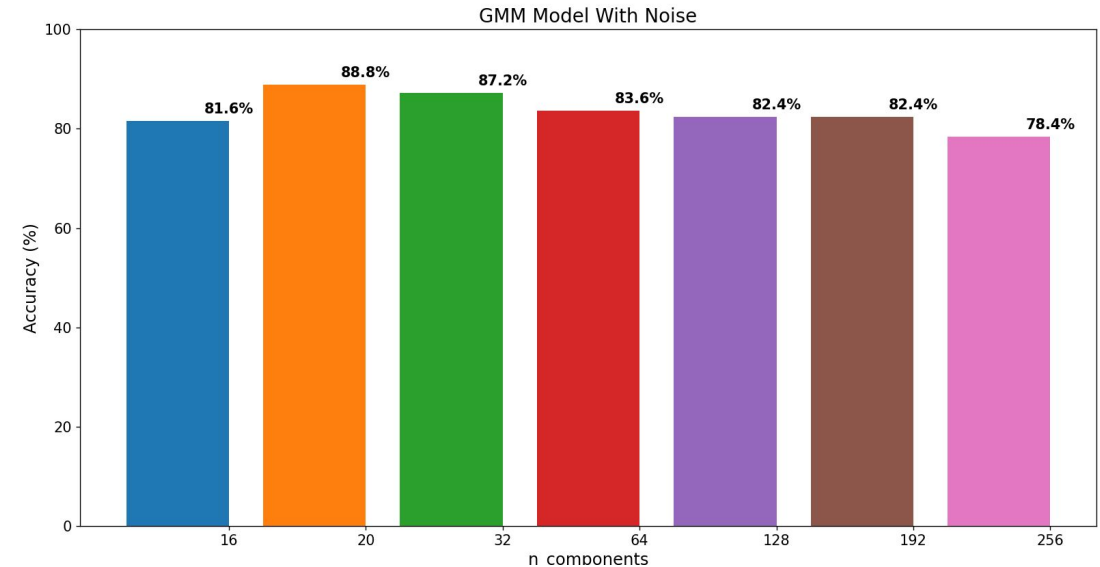
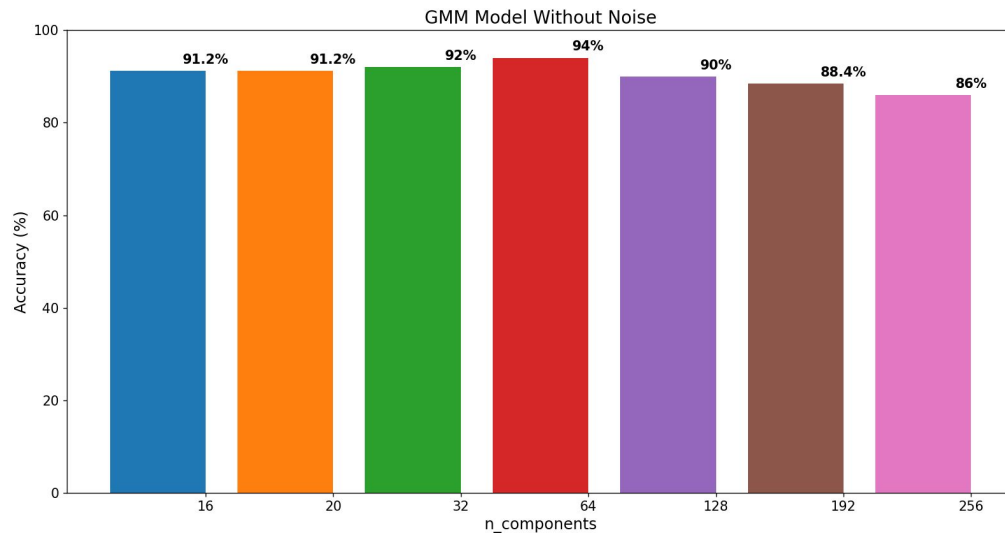
4. Model Training

- Utilizing a Gaussian Mixture Model (GMM) for speaker identification, the model underwent training with varying numbers of components (16, 20, 32, 64, 128, 192, 256) for MFCC, delta, and delta2 features. This comprehensive approach aimed to identify the optimal model configuration for accurate speaker recognition.

Results



- Among the evaluated configurations, training the Gaussian Mixture Model (GMM) **with Noise** with **n_components = 20** yielded the highest accuracy, achieving an impressive **88.8%**.



- Among the evaluated configurations, training the Gaussian Mixture Model (GMM) **without Noise** with **n_components = 64** yielded the highest accuracy, achieving an impressive **94%**.

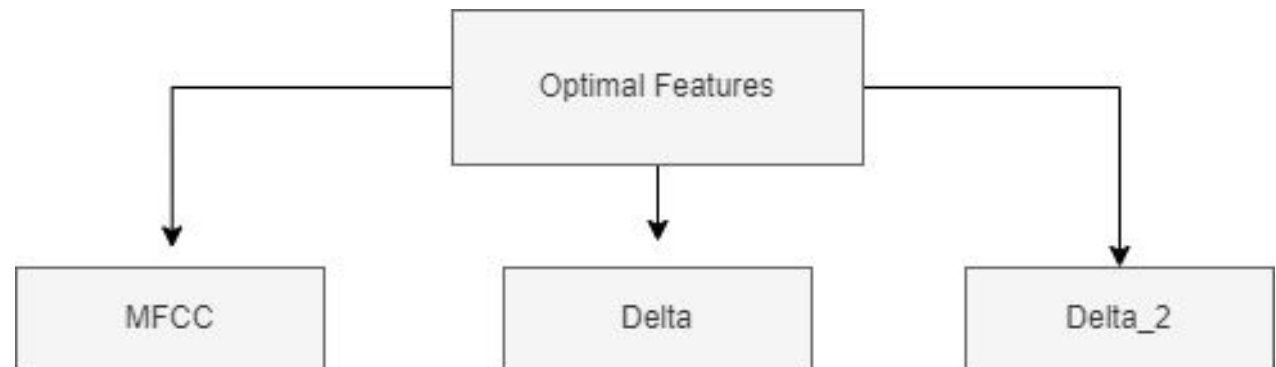
Conclusion



- The experimentation revealed that despite exploring various features, including RMS, ZCR variance, ZCR mean, log energy, delta log energy, LPC, PLP, MFCC, delta, and delta2, the trio of **MFCC, delta, and delta2** emerged as the most effective for achieving high accuracy in speaker recognition.
- These features demonstrated their sufficiency in capturing distinctive speaker characteristics, underscoring their robustness and practicality in the task, simplifying the computational complexity without compromising accuracy.



SPEECH RECOGNITION



THANK YOU

