# Speaker Verification

**K Gopi Krishna**
MT23119

**Mehak Bansal**
MT22111

**Mayank Gupta**
2021065

**Pulkit Nargotra**
2021273

**Vikrant Dhanwadia**
2020263

## Abstract

Speaker recognition, a crucial aspect of biometrics, focuses on the identification and verification of individuals based on their unique vocal characteristics. This field encompasses two primary tasks: speaker identification and speaker verification. In the context of speaker identification, the objective is to ascertain which voice from a predetermined group of known voices best corresponds to the speaker in question. The research aims to delve into the fundamental principles underlying speaker recognition, such as Gaussian Mixture Models (GMMs) and likelihood ratio-based scoring methods. This involves a detailed examination of the theoretical foundations and principles that drive speaker recognition systems. The primary objective of a speaker verification system is to ascertain the authenticity of a speaker's identity through their voice. It delivers a binary outcome: either accepting the identity claim when the acoustic features align with expectations or rejecting it when there's a notable disparity. This technology is instrumental in bolstering security, access control, and identity verification across diverse applications.

## 1 Introduction

Speaker recognition involves identifying individuals by their voice and is categorized into two primary aspects: speaker identification and speaker verification. Speaker identification aims to determine which known voice best matches the speaker, while speaker verification focuses on accepting or rejecting a speaker's identity claim based on acoustic samples. Speaker verification systems are computationally simpler than identification systems, involving a comparison between one or two models instead of one model against multiple speaker models.

We have text-independent speaker. In contrast, text-independent systems, lacking prior text knowledge and user cooperation, require longer utterances for model training and performance reliability.

In this project, our speaker verification system will adhere to the block diagram mentioned above. Within this diagram, the initial stage involves data pre-processing, often referred to as front-end processing. This critical step lays the foundation for subsequent processes in our speaker verification pipeline.

The dataset contains speeches of five prominent leaders namely; Benjamin Netanyau, Jens Stoltenberg, Julia Gillard, Margaret Tarcher and Nelson Mandela which also represents the folder names. Each audio in the folder is a one-second 16000 sample rate PCM encoded. If we combine the chunked audios from 0.wav to 1500.wav, it forms a complete speech of the respective speaker.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

# 2 Literature Review

## 2.1 Intuitive understanding of MFCCs

The study of audio signals and their various features is pivotal in music analysis and speech processing. Among these features, Mel Frequency Cepstral Coefficients (MFCCs) have been extensively employed to extract essential characteristics from audio signals. Typically, a compact set of 10 to 20 features, MFCCs provide valuable insights into the spectral envelope of audio data. Historically, MFCCs have been integral to audio signal processing, particularly in the realm of voice recognition, as exemplified by Muda et al. (2010). In this context, MFCCs serve as fundamental tools for characterizing the spectral properties of audio signals, facilitating the identification of distinctive vocal traits.

Beyond voice recognition, researchers have ventured into the application of MFCCs to describe the intricate concept of "timbre" in music. "Timbre," recognized as the unique quality of sound distinguishing it from others, is inherently multi-dimensional, as highlighted by Peeters et al. (2011). Mauch et al. (2015) embarked on a study aimed at unraveling the evolution of popular music, with a remarkable finding that MFCCs could serve as descriptors for "timbre." This research has raised intriguing questions about the extent to which MFCCs can genuinely capture the complex and multidimensional nature of "timbre" in music.

An essential question arising from the use of MFCCs to describe "timbre" is the interpretability of these coefficients. To understand and validate the conclusions drawn by Mauch et al. (2015), it becomes crucial to provide an intuitive interpretation of MFCCs. These interpretations should shed light on the extent to which these coefficients reflect the nuances and intricacies of "timbre." As the field of audio analysis continues to evolve, with MFCCs playing a central role, the interpretation of these coefficients and their ability to encapsulate the multifaceted nature of "timbre" remain subjects of significant interest and investigation.

In conclusion, Mel Frequency Cepstral Coefficients (MFCCs) have played a pivotal role in audio signal processing, including applications in voice recognition and "timbre" analysis in music. As the field of audio analysis continues to evolve, the interpretation of MFCCs and their capacity to describe the complex nature of audio signals and music remain subjects of significant interest. Future research endeavors may hold the key to further unlocking the potential of MFCCs in understanding the intricate nuances of audio signals and music.

## 2.2 Comparsion with Existing Analysis

Gaussian Mixture Models (GMM) with Mel Frequency Cepstral Coefficients (MFCC), delta, and delta2 have proven to be a robust and well-established method for text-independent speaker recognition. In scenarios where there is no prior knowledge of the spoken content, GMMs have emerged as the most successful likelihood function.

In contrast, text-dependent speaker recognition scenarios benefit from incorporating additional temporal knowledge using hidden Markov models (HMMs) as likelihood functions. In these cases, where there is a strong prior knowledge of the spoken text, HMMs can enhance the modeling of temporal dependencies in speech, potentially improving performance

Likelihood Ratio Interpretation: Utilizes LPCC, LPC, and MFCC features to calculate the probability ratio of aural differences between questioned and sample voices. I-vector Technique: Condenses GMM supervector means into a low-dimensional form, vulnerable to voice disguise, VoIP, and channel mismatch effects.

Support Vector Machine (SVM): Employs SVM classifiers for pattern recognition in speaker verification, demonstrating reported recognition rates. Neural Networks and Deep Learning: Encompasses architectures like MLP, CNN, and DNN, featuring the x-vector for deep speaker embedding and achieving low Equal Error Rates (EER) in experiments.

Table 1: Separate the speech and non-speech segments: using ZCR threshold with Noise

| Speaker | ZCR variance threshold |
|---|---|
| Benjamin Netanyahu | 425 |
| Nelson Mandela | 650 |
| Margaret Tarcher | 300 |
| Julia Gillard | 700 |
| Jens Stoltenberg | 500 |

Table 2: Separate the speech and non-speech segments: using ZCR threshold without Noise

| Speaker | ZCR variance threshold |
|---|---|
| Benjamin Netanyahu | 380 |
| Nelson Mandela | 1000 |
| Margaret Tarcher | 150 |
| Julia Gillard | 380 |
| Jens Stoltenberg | 380 |

## 3 Exploratory data analysis and Preprocessing

### 3.1 Data pre-processing

The initial step in data preprocessing involved the segmentation of background noise into one-second chunks. This segmented noise was then combined with the speech audio through additive mixing. The background noise was sourced from two separate folders containing audio files of varying lengths. From these original audio files, one-second noise clips were created while maintaining the same sampling rate as that of the speech files.

The process ensured that the background noise, extracted from its original sources, was adapted into uniform one-second segments. These segments were subsequently integrated with the speech audio using an additive mixing approach. This methodology aimed to create a dataset where the background noise was standardized and uniformly applied across the speech files, facilitating a consistent experimental environment for subsequent analyses and evaluations.

### 3.2 Removal of non-speech audio files

Zero Crossing Rate (ZCR) surpasses Root Mean Square (RMS) in noise preprocessing for accurate removal of non-speech content. ZCR, measuring the frequency of signal sign changes, excels in capturing rapid variations indicative of speech boundaries. Its heightened sensitivity ensures precise identification of silence intervals, as fewer zero crossings characterize non-speech segments.

Unlike RMS, which gauges overall signal energy, ZCR's focus on dynamic changes enhances its effectiveness in discerning speech from noise. In scenarios with background noise, ZCR remains robust, as it adeptly identifies changes introduced by noise, contributing to the creation of cleaner datasets for subsequent speaker recognition. The nuanced capabilities of ZCR make it a superior choice for enhancing the quality of audio datasets and, consequently, the accuracy of speaker recognition models.

## 4 Methodology and Feature Extraction

### 4.1 Feature Extraction

Feature extraction encompassed various techniques such as MFCC (Mel Frequency Cepstral Coefficients), delta, delta2, ZCR (Zero Crossing Rate) mean and variance, RMS (Root Mean Square) energy, scaling, LPC (Linear Predictive Coding), PLP (Perceptual Linear Prediction), log energy, and delta log energy from audio files. Among these, MFCC, delta, and delta2 demonstrated superior effectiveness compared to other features. ZCR mean and variance showed limited efficacy, while RMS, scaling, LPC, and PLP led to reduced accuracy, emphasizing the prominence of MFCC-based features.

MFCC, delta, and delta2 are highly regarded due to their ability to capture essential speech characteristics effectively. MFCC focuses on the frequency domain and human auditory perception, enabling the representation of speech in a compact form while preserving relevant information. Delta and delta2, representing the temporal derivatives of MFCC, enhance the feature set by encapsulating changes between successive frames, contributing to better discrimination between speech segments.

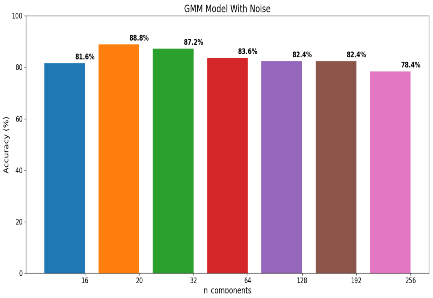
Normalization via dBFS (decibels relative to full scale) is crucial to maintain consistent loudness levels across diverse audio samples. Adjusting each file to a standardized target dBFS level, like -16 dBFS, ensures uniform loudness representation. This normalization process aids in creating a consistent and comparable audio feature set, pivotal for accurate analyses in speaker recognition applications, where consistent representation of features is essential for robust and reliable identification of speakers.

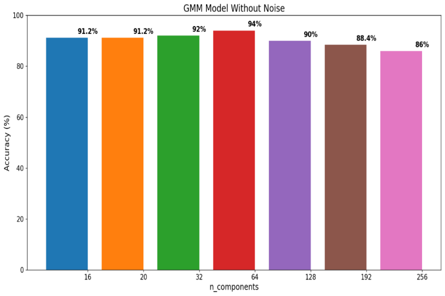
### 4.2 Model Training

The implementation of a Gaussian Mixture Model (GMM) for speaker identification involved a meticulous exploration of model configurations by varying the number of components (16, 20, 32, 64, 128, 192, 256) while limiting the maximum iteration to 10. This iterative approach aimed to pinpoint the optimal model complexity that ensures accurate speaker recognition. The varying components allowed for a nuanced understanding of how the model's discriminative power evolves with increasing complexity, striking a balance between precision and computational efficiency.

The imposed maximum iteration constraint prevented overfitting and contributed to model generalization. This systematic investigation showcases a commitment to refining the GMM for optimal accuracy, offering insights into the nuanced interplay between model parameters and the intricacies of speaker identification, thereby contributing to the advancement of reliable and effective speaker recognition systems.

## 5 Result





In our tests with different setups, we found that training the speaker identification model (Gaussian Mixture Model or GMM) with some background noise, using number of components = 20, gave us really good accuracy—88.8 %. This suggests that considering real-world noise is crucial for accurate results.

On a different note, when we trained the GMM without any background noise and used number of components = 64, we achieved even higher accuracy—94 %. This shows that the best setup can change depending on whether there's background noise or not. So, tweaking the GMM parameters to match the specific conditions of where it will be used is important for getting the most accurate speaker identification.

## 6 Conclusion

In conclusion, thorough experimentation encompassing diverse feature sets—RMS, ZCR variance, ZCR mean, log energy, delta log energy, LPC, PLP, MFCC, delta, and delta2—highlighted the preeminence of MFCC, delta, and delta2 in speaker recognition tasks. This trio exhibited superior effectiveness in capturing unique speaker traits, emphasizing their robustness and practicality.

Their ability to achieve high accuracy underscored their sufficiency in characterizing speakers while simplifying computational complexity without compromising precision. Thus, the conclusive findings endorse the preference for MFCC, delta, and delta2 as the optimal feature set for speaker recognition applications, affirming their pivotal role in reliable and efficient speaker identification systems.

## 7 References

[1] F. Bimbot et al., “A Tutorial on Text-Independent Speaker Verification,” EURASIP J. Adv. Signal Process., vol. 2004, no. 4, p. 101962, Dec. 2004, doi: 10.1155/S1110865704310024.

[2] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication, vol. 17, no. 1–2, pp. 91–108, Aug. 1995, doi: 10.1016/0167-6393(95)00009-D.

[3] J. H. Gambhir and V. V. Patil, “A Review On Speech Authentication And Speaker Verification Methods,” in 2021 Fourth International Conference on Microelectronics, Signals Systems (ICMSS), Kollam, India: IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/ICMSS53060.2021.9673603.