REPORT
SPEAKER IDENTIFICATION REPORT

INTRODUCTION

Speaker identification is a speech processing process that identifies who the speaker is, based on the unique characteristics of their voice. This is a problem of pattern analysis and recognition in a speaker's voice, often through features like pitch, tone, frequency, and pronunciation style. The speaker identification task can be broadly divided into two major tasks: closed-set identification, where the speaker is chosen from a pre-defined set, and open-set identification, where the system determines whether the speaker belongs to an unknown set. This task is critical for applications in security systems, personalized virtual assistants, call center automation, and forensic investigations.

Its importance lies in enabling systems to authenticate individuals, enhancing user experience through tailored responses, and providing reliable evidence in legal scenarios. As voice becomes an increasingly utilized interface, speaker identification serves as a cornerstone for maintaining security, improving user interaction, and supporting technological advancements. It is used across industries for improved security and user experience. Its application in security is in biometric authentication in access control systems for unlocking devices or securing areas. Virtual assistants like Siri or Alexa apply speaker identification to differentiate between household users and to give individual responses. This technology is utilized in call centers to authenticate callers and prevent fraud, thereby streamlining support in customer service. Speaker identification helps the legal and forensic domains analyze audio evidence to verify identities in criminal investigations. It also plays a role in voice-controlled IoT devices, smart homes, and multimedia indexing, enabling seamless interaction and efficient content management.

DATASET

Link to dataset:
https://drive.google.com/drive/folders/1L1NESNH3eqYvbOZeygmd37L6phi6elaL?usp=sharing

The kaggle dataset recommended for us was a dataset with large number of files and the audio duration was very less(1 second). Also few audio files in the recommended dataset were just background noises. So, the dataset for speaker identification is collected from audio recordings from diverse individuals under varying conditions. The data is collected from outside source and the link to the dataset is given. Speech material consists of pre-defined scripts, common phrases, or spontaneous conversations to capture variability in voice patterns. The structured approach ensures a comprehensive dataset suitable for building reliable speaker identification systems.

The dataset used for this speaker recognition project consists of audio files (.wav format) collected from a predefined set of speakers. The audio files were sampled at a standard rate of 16 kHz and saved in mono-channel format. The utterances cover a range of words and phrases, and in total, the dataset comprises recordings from **3 speakers**, with approximately **10 to 15 utterances per speaker**. Each utterance has an average duration of **3–5 seconds**, making the total dataset duration approximately **1–1.5 minutes per speaker**. This ensures that each speaker's voice is adequately represented in the feature extraction and training process.

Dataset Statistics: The following table provides a summary of the dataset's statistics:

Speaker Name	Number of Utterances	Total Duration (Seconds)
Speaker 1	15	60 – 75
Speaker 2	10	30 – 50
Speaker 3	15	60 – 75

(Link for dataset and trained model are given in the README file)

EXPERIMENTAL SETUP

Here is the experimental setup required to efficiently design a speech recognition model:

The process for feature extraction in a speaker recognition system is carried out starting from an audio file. For such a purpose, a particular input directory needs to have audio files arranged so that each audio file exists inside a unique folder belonging to a speaker. Audio processing is then done with regard to Mel-Frequency Cepstral Coefficients (MFCC), segmenting the audio into frames, 25ms with the sampling rate of 16,000 samples per second. MFCC features are extracted for every frame, and afterwards CMS normalization is applied. The features delta are also computed in order to capture temporal derivatives. These features are concatenated together into 40-dimensional feature vectors, and these vectors are saved into the output directory in the format.npy, making it easy to load later steps.

Model training The system constructs a different recognition model for every speaker using Gaussian Mixture Models (GMMs). Users specify the directory containing the extracted feature files, and the system aggregates all feature vectors for each speaker into a single dataset. A GMM with 16 components is trained on these combined features to capture the statistical representation of the speaker's voice. Once the training is done, GMM models are saved in a defined directory, such as gmm_files, having.gmm extension to prepare the models for the test stage.

Testing is the last stage. The system classifies the speaker by choosing the speaker from a recorded audio sample. In the testing stage, a user must specify the path for trained GMM files and input the directory of a test audio file. The same MFCC, CMS, and delta methods used in the feature extraction phase are applied to extract features from the test audio. The system then compares the test features with each of the trained GMM models, scoring them based on likelihood. The model that yields the highest likelihood is used to identify the speaker, and the result is shown to the user.

Result:

The results of the speaker recognition experiment are summarized below, focusing on feature extraction, model training, and testing performance. The analysis highlights the comparative effectiveness of the chosen methods. The use of **MFCC features** with **delta calculations** produced robust 40-dimensional feature vectors, enabling effective speaker characterization. The inclusion of deltas enhanced the model, contributing to better recognition. Using **Cepstral Mean Subtraction (CMS)** and **scaling**, further improved the model's ability to generalize across varied recordings.

The **16-component Gaussian Mixture Model (GMM)** was trained for each speaker using aggregated feature vectors. The number of components can be taken higher or lower but taking lower components shall lead to underperformance and more components lead to good performance but it required very much training time. The GMM with 16 components was selected for its practical balance of accuracy and efficiency.
In the testing phase, the recognition accuracy for known speakers was **90%(**the model correctly recognized the speaker in 9 out of 10 times**)**, with occasional misclassifications due to noisy samples or overlapping vocal characteristics.

Table: showing the efficiency of model for varying values of utterences and number of components

Dataset Size (Utterances)	**Number of Components in GMM**	**Estimated Processing Time (Seconds)**
10	8	5
10	16	10
10	32	20
15	8	7
15	16	15
15	32	30

Conclusion

In this project, we successfully implemented a speaker identification system using MFCC features and Gaussian Mixture Models (GMM) to achieve robust and accurate speaker recognition. The system demonstrated a high recognition accuracy of 90%, leveraging 40-dimensional feature vectors and delta features for effective characterization. The use of 16-component GMMs provided a practical balance between accuracy and computational efficiency, making the solution suitable for real-time applications. Future enhancements could include integrating deep learning-based approaches, such as convolutional neural networks or recurrent neural networks, to improve accuracy further, especially in noisy environments or for overlapping voices. Additionally, expanding the dataset with diverse and multilingual recordings could enhance the system's generalizability and make it adaptable for global applications.

DONE BY(from HITAM COLLEGE):
B.V. DATHATREYA - 22E51A6604
M. PAVAN RAJ - 22E51A6634
S. MIHIR MUDIRAJ - 22E51A6749