Speech Recognition-Based Safety System

Shubham Ram Dabas Kartik Prasad Rahul

Abstract

As digital systems continue to permeate every aspect of our lives, safeguarding sensitive information, such as bank credentials, healthcare records, and personal communications, is more critical than ever. Traditional authentication methods like PINs and passwords are increasingly vulnerable to breaches through hacking, social engineering, or theft. Biometric authentication emerges as a more secure alternative, as it leverages unique physiological and behavioral characteristics of individuals. This project focuses on the development of a text-independent speaker verification system, which authenticates individuals based on their voice rather than specific spoken content. Leveraging features like pitch, tone, and spectral properties, voice authentication offers a contactless, non-intrusive, and secure means of identification. By employing the Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach and fine-tuning it on a specific speaker, we aim to create a robust and adaptable safety system suitable for real-world applications.For more details and access to the code, visit the GitHub repository: https://github.com/Ram21275/ML_Project.git.

1. Introduction

The proliferation of voice-activated systems in devices and applications has brought forth the need for precise and secure identity verification through speech. While identification systems recognize individuals from a pool of known speakers, verification systems confirm a speaker's claimed identity. Traditional methods like PINs and passwords are often prone to security vulnerabilities and lack the convenience that voice-based authentication offers.

Voice serves as an ideal biometric marker due to its distinct attributes shaped by an individual's physical vocal cords and behavioral traits, making it inherently unique. Unlike fingerprints, voice authentication can be performed contactlessly, which is advantageous in scenarios like healthcare or bio-chemical labs, where physical interaction may not be feasible. Additionally, it avoids the intrusive nature of methods like iris scanning. This project aims to build a text-independent speaker verification system that leverages the uniqueness of each individual's voice to provide secure, flexible, and reliable authentication.

2. Literature Review

2.1. Support Vector Machines (SVM) for Speaker Verification

Support Vector Machines (SVM) are powerful and widely used for speaker verification tasks due to their effectiveness in distinguishing between a genuine speaker and an imposter. SVMs operate by constructing a hyperplane that separates two classes—genuine and imposter—based on extracted voice features. They utilize kernel functions, such as the Exponential Radial Basis Function (RBF), to map voice features into a higher-dimensional space, capturing subtle, non-linear patterns within the data.

In this project, a *soft-margin SVM* is employed, allowing for some misclassifications while maximizing the margin between classes. The regularization parameter helps balance margin maximization and classification error minimization, making it suitable for real-world, noisy environments. The use of the RBF kernel enhances the system's ability to distinguish between speakers, focusing on identifying the speaker's identity instead of the spoken content.

2.2. GMM-UBM for Text-Independent Speaker Verification

The *GMM-UBM (Gaussian Mixture Model - Universal Background Model)* approach is a dominant method in speaker verification, particularly for text-independent systems. In this framework, a speaker's voice is modeled as a combination of multiple multivariate Gaussian distributions. The GMM parameters, such as mean vectors, covariance matrices, and mixture weights, characterize the statistical distribution of the speaker's voice features.

The Universal Background Model (UBM) is a general GMM trained on feature vectors from various speakers, capturing common voice characteristics. During training, a speaker-specific GMM is derived by adapting the UBM using *Maximum A Posteriori (MAP) adaptation*, updating mean vectors to reflect speaker-specific traits. Verification is performed by comparing the likelihood ratio between a test sample and both the speaker's GMM and the UBM. If

the likelihood exceeds a threshold, the speaker is authenticated. This method effectively balances general voice traits with individual speaker characteristics, making it robust in real-world applications.

3. Dataset Description

3.1. Attributes and Visualization

The dataset was created by recording each speaker for 15-20 minutes, segmented into 2-second intervals, resulting in numerous voice segments. These segments capture various vocal traits and environmental conditions, adding diversity to the dataset.

The primary features extracted are *Mel Frequency Cepstral Coefficients (MFCCs)*, which capture the spectral characteristics of the speaker's voice. Additional features such as pitch, tone, and spectral flatness were also analyzed.

For visual analysis, spectrograms were generated to show frequency changes over time, helping identify distinct speaker patterns. Histograms of MFCC distributions further highlight vocal tract and tonal variations across speakers.

3.2. Preprocessing

Voice data preprocessing was crucial for improving the quality of the recorded samples. The primary challenge in speaker verification is isolating active speech segments while filtering out silent or non-speech portions. To address this, we applied *Voice Activity Detection (VAD)*, which identifies and removes periods of silence, ensuring that only relevant voice segments are used for model training and testing. This technique is especially beneficial in scenarios with background noise or when speakers pause during their utterances.

3.3. Fine-Tuning with Ram Dabas' Voice

In this project, the dataset was fine-tuned using Ram Dabas' voice samples to serve as the reference for speaker verification. After the initial training on the full dataset, the *Universal Background Model (UBM)* was adapted specifically to Ram's voice through *Maximum A Posteriori (MAP) adaptation*. This adaptation updated the UBM's parameters to better reflect Ram's vocal characteristics, creating a more accurate speaker-specific model. This personalized model for Ram Dabas allowed the system to verify his identity with higher accuracy compared to other speakers.

By leveraging both the dataset and Ram's fine-tuned model, we aimed to achieve a robust and reliable speaker verification system capable of handling diverse content and variable conditions.

4. Methodology and Model Details

4.1. Dataset Preparation with Background Noise

To prepare the dataset for training the model with background noise, we followed a systematic approach involving the following steps:

1. **Segmenting Audio Files:** We first took **2-second audio segments** from the voice samples. These segments were taken from the voice of the original speaker (in this case, Ram) as well as from imposter voices(rest of the group members).

2. **Adding Background Noise:** To simulate real-world conditions where background noise is present, we selected **10 noise files** containing typical environmental sounds. We then mixed these noise files with the clean speech segments to create noisy audio samples.

3. **Normalization and Mixing:** Both the speech files and noise files were **normalized** to ensure that the volume levels were consistent across all the files. After normalization, we **merged the noise files** with the clean speech segments at different **mixing ratios**. Specifically, we experimented with the following five mixing ratios:

    - 30% noise, 70% speech
    - 40% noise, 60% speech
    - 50% noise, 50% speech
    - 60% noise, 40% speech
    - 70% noise, 30% speech

4. **Combining the Noisy Dataset:** After applying each of the mixing ratios to the speech and noise, we **merged all the noisy samples** (from different ratios) into a single dataset. This process was repeated for both the voice of the original speaker (Ram) and the imposter voices.

5. **Dataset Size and Training Configuration:** The resulting dataset was large enough to train the model for robust speaker recognition in noisy environments. The training was conducted over **100 epochs** with a batch size of **32** samples, which allowed the model to learn the variations in both clean and noisy speech.

By following this procedure, we were able to generate a diverse dataset of speech samples with varying noise levels, helping the model to generalize better when dealing with noisy real-world audio inputs.

4.2. Feature Extraction (MFCC)

To accurately capture the unique vocal characteristics of each speaker, we employed *Mel Frequency Cepstral Coefficients (MFCCs)* as the primary feature representation. MFCCs are well-suited for speaker verification tasks as they effectively model the vocal tract characteristics by analyzing short-term power spectra of speech signals. Each audio segment was processed to extract a set of MFCCs, capturing key properties like formants, pitch, and spectral energy distribution. The coefficients were computed using a window size of 25 ms and a step size of 10 ms to ensure detailed spectral analysis.

In addition to MFCCs, we considered other features, such as spectral flatness and zero-crossing rate, to further explore their impact on speaker distinction. However, MFCCs demonstrated superior effectiveness in retaining speaker-specific information.

4.3. Preprocessing Techniques

Cepstral Mean Subtraction (CMS): CMS reduces channel noise by subtracting the mean cepstral coefficients from each frame, normalizing the speech signal and highlighting speaker-specific features.

Band-Pass Filtering: Band-pass filtering isolates the speech frequency range (80 Hz to 3000 Hz and 600 Hz to 3000 Hz), removing low- and high-frequency noise, enhancing the speech signal for better speaker verification. Got better audio files on 80 Hz - 3000 Hz.

Wiener Filtering: Wiener filtering reduces noise by adapting to local noise characteristics and minimizing the mean square error between the noisy and clean signals, effectively preserving speech in noisy environments.

4.4. Universal Background Model (UBM) Training

The next step involved training a *Universal Background Model (UBM)* on the extracted MFCCs from all speakers in the dataset. The UBM, modeled as a *Gaussian Mixture Model (GMM)*, serves as a baseline for capturing general voice characteristics that are speaker-independent. We utilized a mixture of 64 Gaussian components to represent the various spectral patterns found across speakers.

The UBM training process involved clustering the MFCCs from all speakers using the Expectation-Maximization (EM) algorithm. This approach allowed us to derive the mean vectors, covariance matrices, and mixture weights for each Gaussian component, representing the overall voice space in a speaker-independent manner.

4.5. Speaker-Specific Model Adaptation

For each speaker, the UBM was adapted to create a speaker-specific GMM using the *Maximum A Posteriori (MAP) adaptation* technique. During adaptation, the mean vectors of the UBM's Gaussian components were updated based on the speaker's MFCCs, while the covariance matrices and weights were adjusted to fine-tune the model to the individual's voice. This adaptation process allowed the GMM to capture the unique vocal characteristics of each speaker while retaining general voice traits from the UBM.

The MAP adaptation process aimed to balance the amount of adaptation data with the prior information obtained from the UBM. By incorporating only the relevant information from the speaker's data, this approach improved the robustness of the adapted model in real-world scenarios with limited training data.

4.6. Speaker Verification

The speaker verification task was performed by comparing the log-likelihood scores of a test audio sample against both the UBM and the speaker-specific GMM. The system computed the log-likelihood ratio between these scores to determine the likelihood of the test sample belonging to the claimed speaker. A predefined threshold was set to make the authentication decision: if the log-likelihood ratio exceeded the threshold, the speaker was accepted; otherwise, the speaker's identity was rejected.

To further enhance the robustness of the system, multiple scores from consecutive audio segments were averaged over a sliding time window. This averaging helped mitigate the impact of temporary variations in pronunciation or background noise, leading to more consistent verification results.

4.7. Model Selection (SVM Testing)

To explore alternative classification methods, we tested a *Support Vector Machine (SVM)* classifier with a soft-margin configuration. The SVM utilized an *Exponential Radial Basis Function (RBF)* kernel to effectively handle the non-linearities present in the voice data. The RBF kernel mapped the MFCC features into a higher-dimensional space, allowing the SVM to capture subtle variations in voice characteristics.

During SVM training, we performed grid search-based hyperparameter tuning to identify the optimal values for the regularization parameter and the RBF kernel coefficient. The SVM's performance was evaluated in comparison to the GMM-based approach, focusing on its ability to differentiate genuine speakers from imposters.

5. Results and Analysis

In this study, we implemented and evaluated a speaker authentication system using Gaussian Mixture Models (GMM) with Mel Frequency Cepstral Coefficients (MFCC) for feature extraction. Additionally, we tested a Support Vector Machine (SVM) classifier as an alternative during

the model selection process. The performance of both models was assessed using a dataset of voice recordings from multiple speakers.

5.1. Speaker Verification with GMM-UBM and Noisy Data

The performance of the GMM-UBM model under various noise conditions and preprocessing techniques is summarized as follows:

No Preprocessing: On noisy data without preprocessing, the GMM-UBM achieved an accuracy of 73.68% and an Equal Error Rate (EER) of 25.9%, serving as the baseline for comparison.

CMS (Cepstral Mean Subtraction): Applying CMS resulted in a slight accuracy decrease 73.19% and a higher EER 29.61%. While CMS normalizes channel effects, it is less effective in handling the complex noise distortions in this dataset.

Wiener Filtering: Wiener filtering reduced the accuracy to around 63% with an EER of 32%. It attenuates noise but may distort speaker characteristics, affecting verification performance.

Bandpass Filtering: Bandpass filtering achieved an accuracy of about 61% and an EER of 33%, indicating that while it cleans the signal, it may also remove important speaker-specific information.

Clean Data: When trained and evaluated on clean data, the model achieved an accuracy of 91.96% and an EER of 1.02%, highlighting the impact of noise on verification performance.

5.2. Conclusion

5.3. SVM Model Performance

We tested an SVM classifier on the dataset, achieving an accuracy of *74.566%* on the clean (noise-free) data. While this performance indicates that the model can distinguish between speakers, it was still significantly lower than the GMM-based system.

5.3.1 Performance with Noisy Data

When applied to noisy data, the SVM's accuracy dropped to *61.54%* with an Equal Error Rate (EER) of *38.46%*. Even after applying noise reduction techniques like **Wiener Filtering**, the accuracy remained almost unchanged, suggesting that the SVM was not robust enough to handle noise effectively.

5.4. Comparative Analysis

The GMM model outperformed the SVM model in both accuracy and reliability. The significantly lower EER for the GMM-based approach further emphasizes its robustness in speaker authentication, making it the preferred choice for this task. The performance gap between the GMM and SVM can likely be attributed to the nature of voice data, where GMMs are particularly well-suited for modeling the probabilistic distributions of speaker-specific features.

6. Conclusion

The GMM-UBM model outperformed the SVM classifier in speaker authentication, achieving **91.96% accuracy** with an **EER of 1.02%** on clean data. In noisy conditions, GMM-UBM achieved **73.68% accuracy** with an **EER of 25.9%**, surpassing the SVM, which achieved only **61.54% accuracy**. Preprocessing methods such as CMS, Wiener filtering, and bandpass filtering showed limited improvements, emphasizing the challenges of handling noise. GMM-UBM's superior performance, especially in clean conditions, demonstrates its suitability for speaker verification, with potential for further enhancement through advanced noise-robust features and fine-tuning techniques like MAP adaptation.

7. Contributions

The development of our speaker authentication system was the result of collaborative efforts by all team members. The specific contributions were as follows:

- **Literature Review:** Kartik conducted an in-depth review of existing research in the field of speaker authentication.
- **Model Selection:** Shubham and Kartik focused on testing the SVM classifier, while Ram and Rahul worked on testing the GMM-UBM model.
- **Data Visualization:** Shubham handled the task of visualizing the dataset to facilitate analysis and model understanding.
- **Preprocessing and Fine-tuning:** Ram was responsible for preprocessing, voice activity detection (VAD), and fine-tuning using Maximum A Posteriori (MAP) adaptation.
- **Report Preparation:** Shubham, Rahul, Ram, and Kartik collaboratively authored the project report, ensuring comprehensive documentation of methodologies and findings.
- **Presentation Preparation:** Rahul led the preparation of the final project presentation, synthesizing key findings and results.

References

- Speaker Verification Using MFCC and Support Vector Machine

- Noise robust speaker verification using GMM-UBM multi-condition training