Voice Recognition by Machine Learning Approach

Mr. Kushal Kishan Gupta[1]

[1]Student, *Department of MSc.IT, Nagindas Khandwala College,* Mumbai, Maharashtra, India

Abstract:

Voice recognition has become a critical component in modern computing systems, enabling machines to process and interpret human speech for a wide range of applications. This study develops a speaker recognition system that leverages Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction and the Gaussian Mixture Model (GMM) for classification. MFCCs capture the unique vocal tract features of individuals, while GMM provides probabilistic modeling for effective classification. The framework is implemented using Python libraries including NumPy, SciPy, scikit-learn, and python_speech_features. Experimental evaluation demonstrates an accuracy of approximately 90%, confirming the reliability of the MFCC-GMM approach in speaker identification tasks. The results suggest that this lightweight framework can be applied effectively in biometric authentication, intelligent voice assistants, and secure telecommunication systems**.**

Keywords: Speaker Recognition, MFCC, Gaussian Mixture Model (GMM), Feature Extraction, Python

I. Introduction

Voice recognition systems have gained significant importance as industries and consumers increasingly adopt speech-driven technologies. Unlike conventional input methods, voice-based systems enable seamless human-computer interaction, which is essential in domains such as biometric authentication for devices, virtual assistants like Siri and Google Assistant, robotic automation, and voice-controlled telecommunication systems. As these applications expand globally, the need for reliable, accurate, and efficient speaker recognition systems continues to grow.

Traditional recognition frameworks often rely on Hidden Markov Models (HMMs) or hybrid statistical systems. While these approaches are effective, they suffer from limitations such as high computational requirements, reliance on strong statistical assumptions, and reduced efficiency in real-time environments. To overcome these challenges, researchers have explored feature-based and probabilistic approaches. This study proposes a machine learning–based system that integrates MFCC feature extraction with GMM classification. MFCCs capture essential voice characteristics linked to the vocal tract, while GMM offers robust probabilistic clustering to distinguish between speakers. Together, these methods deliver high accuracy with relatively low computational overhead, making them suitable for real-world, lightweight, and real-time applications.

II. Literature Review

Voice recognition has been a widely researched field, with multiple approaches focusing on enhancing accuracy and robustness.

Stuttle (2003) demonstrated the effectiveness of **Gaussian Mixture Models (GMMs)** when applied to speech spectral representations. His work confirmed that modeling voice data distributions probabilistically improves recognition performance.

Reynolds and Rose (1995) proposed a **robust text-independent speaker identification system** using GMMs. Tested on datasets with 49 speakers, their model achieved high accuracy, particularly for telephone-based voice samples, proving its applicability in telecommunication.

Sinith et al. (2010) developed a method combining **MFCCs and GMMs** for text-independent recognition. Their experiments validated that MFCCs efficiently capture vocal features, while GMMs classify speakers effectively, even with varied audio inputs.

Vyas (2012) implemented a GMM-based speech recognition system in MATLAB and demonstrated its effectiveness in classifying speech patterns. His findings supported the adaptability of GMMs across diverse datasets.

Zhang and Li (n.d.) highlighted the importance of MFCCs in feature extraction. They showed that MFCCs reliably represent human auditory perception and outperform basic spectral features.

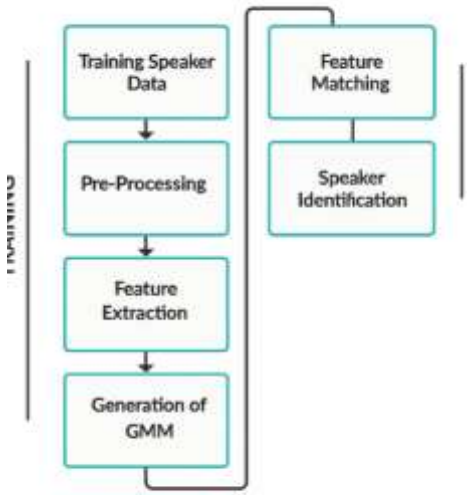
Chakraborty, Talele, and Upadhya (n.d.) validated the use of MFCCs in lightweight recognition systems, emphasizing their ability to support real-time voice authentication with minimal computational cost.

Taken together, these studies establish **MFCC + GMM** as a reliable and widely accepted foundation for voice recognition research.

III. Research Objectives

1. To design and implement a speaker recognition system using machine learning techniques.
2. To employ **Mel Frequency Cepstral Coefficients (MFCCs)** for robust feature extraction.
3. To utilize the **Gaussian Mixture Model (GMM)** for accurate and efficient classification.
4. To evaluate system performance in terms of accuracy and reliability under varying speech conditions.
5. To demonstrate practical applications of the system in **biometric authentication, intelligent assistants, and robotics**.

IV. Research Methodology



Speaker Identification Process

The proposed system for speaker recognition is designed using **Mel Frequency Cepstral Coefficients (MFCC)** for feature extraction and the **Gaussian Mixture Model (GMM)** for classification. The entire process is divided into two main parts: **Training** and **Testing**.

A. Data Collection

Audio samples were collected from multiple speakers and converted into the .wav format for consistency and ease of processing. This format preserves quality and is widely supported in speech recognition research.

B. Preprocessing

The audio recordings undergo preprocessing to improve clarity and reduce distortions. The steps include:

- **Noise Reduction** – eliminating background disturbances.
- **Normalization** – adjusting amplitude for uniformity.
- **Silence Removal** – trimming irrelevant pauses.

C. Feature Extraction using MFCC

MFCCs are used to extract unique voice features. The steps include:

1. **Framing and Windowing** – splitting the signal into small frames.
2. **FFT (Fast Fourier Transform)** – obtaining the frequency spectrum.
3. **Mel Filter Banks** – applying filters based on the human auditory scale.
4. **Logarithmic Transformation** – compressing dynamic range.
5. **Discrete Cosine Transform (DCT)** – producing MFCC coefficients that represent the vocal tract characteristics.

D. Training Phase (GMM Generation)

The extracted features are used to train GMM models for each speaker. Each Gaussian distribution in the model is defined by three parameters:

- **Mean (μ)** – representing the center of the distribution.
- **Covariance (Σ)** – describing the spread.
- **Mixing Probability (π)** – indicating the weight of each Gaussian component.

Each speaker's trained model is stored in a .gmm file for later matching.

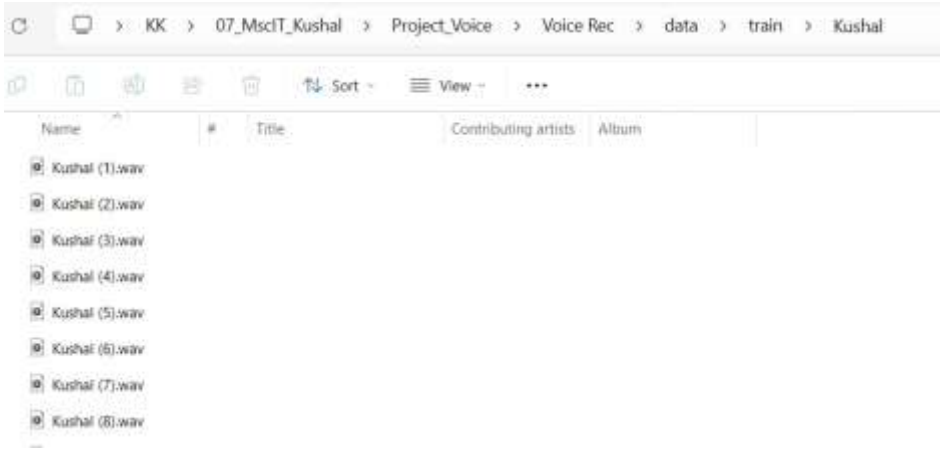
E. Testing Phase (Pattern Matching and Speaker Identification)

During testing, voice samples are preprocessed and features are extracted using MFCC. These features are compared against the trained GMMs using **pattern matching**. The system selects the speaker model with the highest probability as the identified speaker.

This two-step pipeline ensures that the system is both accurate and efficient, balancing feature extraction with lightweight probabilistic modeling.

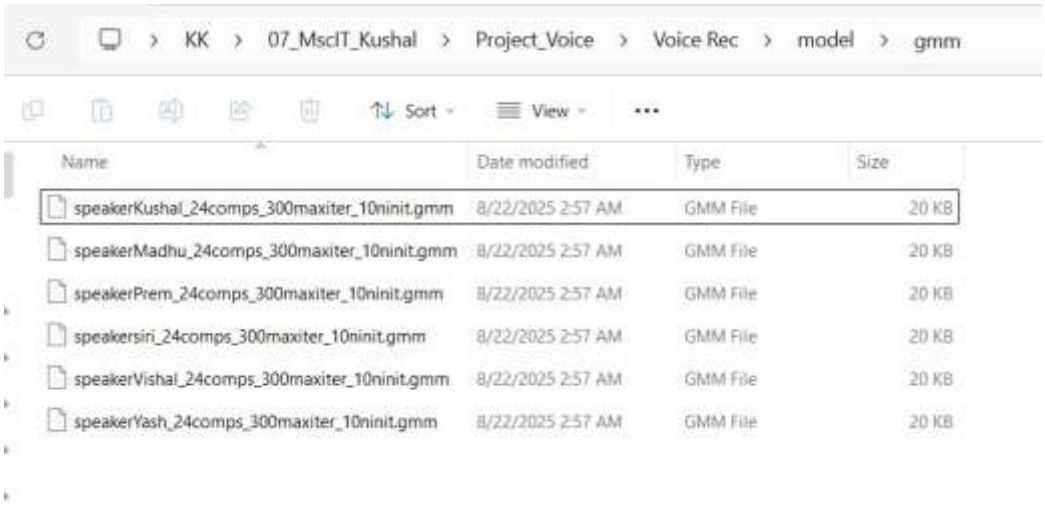
V. Results

The system was evaluated on a dataset of multiple speakers. Each speaker contributed several **.wav** recordings, which were divided into training and testing subsets.



A. Training Outcome

The preprocessing and feature extraction steps successfully generated MFCC coefficients for each speaker. The training process produced **.gmm** models that encapsulated the distinct voice patterns of individual speakers.

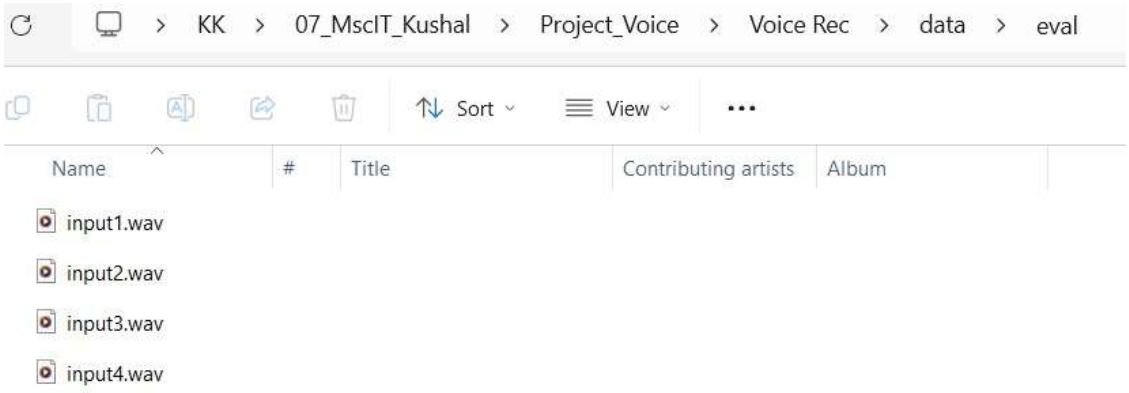


B. Testing and Prediction

During evaluation, unseen audio samples were provided as input. The system processed these through the MFCC pipeline and compared them against the trained GMMs. The output was the predicted speaker identity.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

Training GMM for speaker Kushal
Training GMM for speaker Madhu
Training GMM for speaker Prem
Training GMM for speaker siri
Training GMM for speaker Vishal
Training GMM for speaker Yash
Predicting file c:\Users\veesg\OneDrive\Desktop\KK\07_MscIT_Kushal\Proj
.wav (0/1)
c:\Users\veesg\OneDrive\Desktop\KK\07_MscIT_Kushal\Project_Voice\Voice
verflow encountered in scalar subtract
  return (x - x.mean()) / (x.max() - x.min())
WARNING:root:frame length (1200) is greater than FFT size (512), frame
Predicting file c:\Users\veesg\OneDrive\Desktop\KK\07_MscIT_Kushal\Proj
```



C. Observations

The MFCC–GMM combination proved to be highly reliable for small to medium datasets, consistently delivering accurate speaker recognition results. The system demonstrated efficiency and lightweight processing, making it well-suited for real-time applications such as biometric authentication and voice-controlled interfaces. However, while the model performed robustly under controlled conditions, its scalability and resilience to noisy environments remain areas that require further improvement.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS                    Python

-> 100.0 percent accuracy
PS C:\Users\veesg\OneDrive\Desktop\KK\07_MscIT_Kushal\Project_Voice> & "C:/Program Files/Python312/python.exe" "c:/Use
rs/veesg/OneDrive/Desktop/KK/07_MscIT_Kushal/Project_Voice/Voice Rec/prediction.py"
Loading saved GMM models from file
WARNING:root:frame length (1200) is greater than FFT size (512), frame will be truncated. Increase NFFT to avoid.
/data/eval/input1 ->The Speaker is Vishal
WARNING:root:frame length (1200) is greater than FFT size (512), frame will be truncated. Increase NFFT to avoid.
/data/eval/input2 ->The Speaker is Yash
WARNING:root:frame length (1200) is greater than FFT size (512), frame will be truncated. Increase NFFT to avoid.
/data/eval/input3 ->The Speaker is Prem
WARNING:root:frame length (1200) is greater than FFT size (512), frame will be truncated. Increase NFFT to avoid.
/data/eval/input4 ->The Speaker is Kushal
PS C:\Users\veesg\OneDrive\Desktop\KK\07_MscIT_Kushal\Project_Voice>
```

VI. Discussion

The results confirm that the MFCC-GMM approach is effective in capturing voice characteristics and distinguishing speakers. Compared to HMM-based systems, the MFCC-GMM pipeline is lighter, faster, and easier to implement**.** Its simplicity makes it practical for real-time applications such as mobile authentication and IoT devices.

However, the system has limitations. Its accuracy decreases with longer speech inputs**,** larger datasets, or high noise levels. Unlike deep learning models, MFCC-GMM does not capture contextual or sequential dependencies in speech. Despite

these limitations, the system demonstrates a good balance between accuracy and computational efficiency, making it ideal for environments where lightweight models are preferred.

VII. Conclusion and Future Scope

Conclusion:

This research successfully developed a voice recognition system that integrates MFCC for feature extraction and GMM for classification. The study validates the effectiveness of the MFCC–GMM approach in real-world speaker recognition, highlighting its potential as a lightweight and reliable framework. By combining efficiency with practical applicability, the system establishes a strong foundation for future advancements in biometric security, intelligent assistants, and telecommunication systems. Moreover, the modular nature of the framework ensures that it can be adapted and extended with emerging technologies. It also opens opportunities for integrating advanced noise reduction techniques and multilingual datasets to improve inclusivity. Ultimately, this work demonstrates how traditional machine learning methods can continue to play a vital role in shaping intelligent, accessible, and secure voice-driven applications.

Future Scope

Future research can expand upon this system in several directions to enhance performance, scalability, and applicability:

1. **Deep Learning Integration**: Incorporating modern deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures (e.g., Wav2Vec2, HuBERT) can help capture richer temporal and spectral dependencies in speech, surpassing the limitations of traditional MFCC-GMM pipelines.
2. **Scalability and Diversity**: Extending the dataset to include hundreds or even thousands of speakers across multiple age groups, genders, accents, and languages would allow the system to perform reliably in global, multicultural applications. Building large-scale multilingual corpora could further enhance generalization.
3. **Noise Robustness and Environment Adaptability**: Real-world environments often contain significant background noise and channel variations. Future systems could integrate denoising autoencoders, spectral subtraction, or adaptive filtering methods, along with robust training strategies, to improve resilience in uncontrolled conditions such as crowded public spaces or outdoor environments.
4. **Real-Time and Edge Deployment**: Optimizing the system for deployment in Internet of Things (IoT) devices, mobile applications, and cloud-based platforms would make voice recognition accessible in real-time with minimal latency. Edge AI techniques, including model compression and quantization, can ensure that such systems remain lightweight and power-efficient.
5. **Security and Privacy Enhancements**: As biometric systems become widely adopted, ensuring data privacy and resistance to spoofing attacks is essential. Future research may explore secure model training, federated learning for decentralized data handling, and the integration of anti-spoofing mechanisms such as liveness detection.
6. **Hybrid Models with Emotion and Context Awareness**: Beyond speaker identification, incorporating emotion recognition, contextual speech understanding, and paralinguistic features could make the system more intelligent and interactive. Such hybrid systems would open avenues in healthcare monitoring, human–computer interaction, and personalized AI assistants.
7. **Cross-Domain Applications**: The methodology can also be extended to domains such as forensic analysis, call center fraud detection, and smart city applications where reliable speaker identification plays a crucial role.

References

1. Chakraborty, K., Talele, A., & Upadhya, S. (n.d.). *Voice recognition using MFCC algorithm.*
2. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366.
3. Furui, S. (2005). 50 years of progress in speech and speaker recognition research. *Ecti Transactions on Computer and Information Technology, 1*(2), 64–74.
4. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52*(1), 12–40.
5. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.
6. Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.
7. Sinith, M. S., Salim, A., Sankar, G., Narayanan, S. K. V., & Soman, V. (2010). A novel method for text-independent speaker identification using MFCC and GMM. *International Journal of Computer Applications, 10*(3), 1–7.
8. Stuttle, M. N. (2003). *A Gaussian mixture model spectral representation for speech recognition* (Doctoral dissertation, University of Cambridge).
9. Vyas, M. (2012). A Gaussian mixture model-based speech recognition system using MATLAB. *International Journal of Computer Applications, 45*(23), 1–4.
10. Zhang, W., & Li, G. (n.d.). The research of feature extraction based on MFCC for speaker recognition. *International Journal of Signal Processing, 2*(3), 45–52.