SPEAKER RECOGNITION USING GAUSSIAN MIXTURE MODEL

Ms. Allada Keerthana[(1)],
Assistant Professor, Department of Statistics
St.Ann's College for Women, Mehdipatnam, Hyderabad, Telangana, India.
+91 7396765785

Ms. Kalaga V N S Anjanee Gayatri[(2)],
Assistant Professor, Department of Statistics
St.Ann's College for Women, Mehdipatnam, Hyderabad, Telangana, India.
+91 9515697414, ORCID:0009-0000-3934-5012

Abstract
The research focuses on presenting a speaker identification system using audio recordings and machine learning techniques, specifically Mel-frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs), to recognize individuals based on their unique voice characteristics. The modular pipeline includes data acquisition, feature extraction, model training, and identification. This study aims to validate the classical MFCC-GMM framework for speaker identification in controlled settings. Applications such as secure authentication, fraud detection, and user personalization across industries highlight the growing importance of this technology. While the system demonstrates perfect accuracy under ideal conditions, future expansion is needed for real-world applicability.

Keywords: Speaker Recognition, Speaker Identification, Voice Biometrics, MFCC (Mel-Frequency Cepstral Coefficients), Gaussian Mixture Model (GMM), Speech Processing, Voice Authentication, Pattern Recognition

1. Introduction

A need for reliable and efficient speaker recognition systems has arisen due to the increasing use of voice-driven human-computer interaction. There is an increasing need for strong speaker identification systems that can reliably identify single users based just on their vocal qualities as voice-based technologies grow pervasive from virtual assistants and smart house devices to secure banking and customer service automation. One branch of biometric technologies, speaker identification is the process of confirming or checking the identity of a person using their voice. Compared with more conventional biometric techniques like fingerprints or face recognition, this approach has several benefits. It allows contactless communication, needs only a typical microphone for input, and enables distant authentication in situations where physical identification is impossible or impracticable.

Two basic kinds of speaker recognition are speaker verification and speaker identification. A one-to-one check called speaker verification confirms whether the claimed identity of a speaker corresponds with a previously saved voice model. Conversely, speaker identification is a one-to-many categorization challenge in which the machine must identify an anonymous speaker from among a collection of registered users. The present study concentrates on speaker identification, particularly under text-independent conditions, where there is no restriction on the spoken material during enrollment or recognition. This scenario more closely matches real-world uses and presents more difficulties as the system must generalize over many speech patterns, lexical content, and emotional states.

The human voice has unique qualities affected by behavioral elements including speaking style, accent, pitch as well as anatomical features including the vocal tract, larynx, and nasal cavity. Finding and representing these changes needs a mix of strong feature extraction and adaptable statistical modeling. Among many methods, the usage of Mel-Frequency Cepstral Coefficients (MFCCs) has become a common practice in speech signal processing.(Lyons, 2018). MFCCs allow subsequent machine learning algorithms to quickly learn patterns unique to each speaker by converting

the speech signal into a small series of feature vectors. Furthermore, to statistically simulate the distribution of MFCC characteristics for every speaker Gaussian Mixture Models (GMMs) are used in this study. (Reynolds, Quatieri, & Dunn, 2000). With their probabilistic approach, GMMs portray the underlying feature space as a weighted sum of multivariate Gaussian distributions. This gives the system flexibility in simulating varied vocal patterns while still recording speaker-specific features. The MFCC-GMM framework has shown competitive performance in many restricted situations and acts as a fundamental baseline for complicated deep learning techniques.

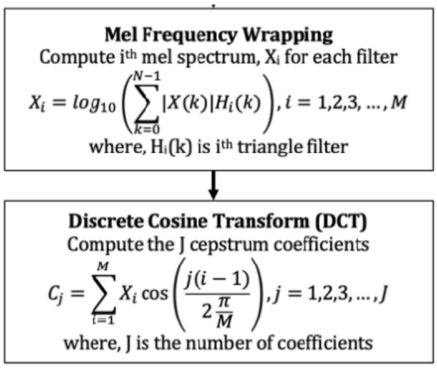

Fig. 1: Probabilistic Formulation of a Gaussian Mixture Model (GMM)

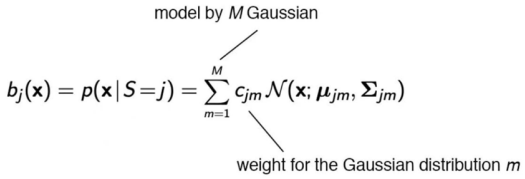

Fig. 2: Mel-Frequency and Cepstral Coefficient Computations

The development of a modular speaker recognition method meant to properly identify people using minimal, high-quality audio data is shown in this study. Validated under regulated circumstances, the framework shows great identification performance and applies across several fields including security, customized computing, and criminal inquiries.Its adaptability and simplicity make it not only a practical solution but also a valuable tool.

2. Literature review

The field of speaker recognition has been a significant area of study in speech processing and biometric authentication, with researchers constantly improving systems to make them more accurate or adaptable. Capturing and identifying speaker-specific qualities from audio signals is the main challenge of this study; it starts with good feature extraction.

The Mel-Frequency Cepstral Coefficient (MFCC) is one of the most often used methods for this objective. MFCCs convert the speech signal into small characteristics that mirror its spectral envelope—a vital component in recognizing a speaker's distinct vocal traits. It is inspired by the way people hear sounds. Particularly in settings where efficiency and interpretability are important, MFCCs have continuously shown better performance than alternatives like LPC, PLP, or formant-based traits.

Equally crucial is the selection of a model to depict these characteristics. Their capacity to model complicated speech patterns has made Gaussian Mixture Models (GMMs) quite the standout choice. (Reynolds et al., 2000). Using several Gaussian components, every GMM represents the statistical distribution of a speaker's MFCCs. Reynolds and others discovered that GMMs, with appropriate training using Expectation-Maximization, can provide great identification accuracy even with little data. Intern researchers have enhanced their models using techniques like Maximum a Posteriori (MAP) adaptation and Universal Background Models (UBMs) to raise performance further. Better generalization from smaller enrollment data sets is made possible by these methods, which have also become usual practices in more advanced systems.

Deep learning has greatly accelerated the field more lately. Architectures like CNNs, RNNs, and attention-based models have allowed for the use of speaker embeddings like i-vectors and x-vectors, which function well in noisy, huge scale settings. (Snyder et al., 2018). Still, they have greater complexity and need of resources. By their simplicity, explainability, and ease of installation, conventional MFCC-GMM systems remain enticing therefore perfect for foundational uses and instruction.

Although not employing contemporary neural methods, the system fits with long-standing research and shows that when applied deliberately and under regulated circumstances, even conventional methods can provide dependable performance. This project builds on that classic base using a simple MFCC-GMM pipeline to prove its ongoing usefulness.

3. Methodology

The speaker identification system implemented in this study is organized by making use of a modular pipeline consisting four fundamental phases: audio data collecting, feature extraction, model training, and speaker prediction, the speaker recognition system used in this research is Every stage has been painstakingly created and progressively connected so that a functional and replicable system for finding personal speakers based on their voice samples may be possible. The approach underlines the use of understandable, widely used methods that strike a balance between simplicity and efficiency, therefore making the system appropriate for both scholarly investigation and possible real-world application in restricted environments.

Starting with the recording of voice samples from every speaker, audio data acquisition starts here. The system enables users to offer their voice using a microphone for a predetermined length—usually 10 seconds per recording. To reduce ambient noise and guarantee clear input signals for this investigation, voice samples from two different speakers were recorded several times in silent indoor settings. The collected samples are kept in uncompressed WAV format to guarantee audio quality and to enable dependable feature extraction in later phases. To facilitate regulated performance assessment of models, the recordings are meticulously labeled and arranged into training and testing sets.

The next stage of the pipeline is extracting significant characteristics from the raw audio waveforms. Because of their proven ability to capture the perceptually pertinent elements of speech, Mel-Frequency Cepstral Coefficients (MFCCs) are used as the main features for this goal. (Lyons, 2018). Beginning with pre-emphasis to boost high-frequency elements and lower signal-to-noise ratio, the feature extraction procedure follows a conventional chain of signal processing stages. The signal is subsequently split into overlapping frames, and each frame is windowed using a Hamming window to reduce spectral leakage. Then a Mel-scaled filter bank is used to imitate the logarithmic pitch perception of the human ear after a Fast Fourier Transform (FFT) is performed to convert each frame into the frequency domain. The ultimate MFCC vectors are generated by calculating the log filter bank energies and applying a Discrete Cosine Transform (DCT). Along with the static MFCC coefficients, the system generates delta and delta-delta coefficients—representing the temporal derivatives of the traits. Captured changes in the speech signal across time, these dynamic characteristics have been shown to increase classification performance in speaker identification activities.

After feature extraction, the system goes on to the model training stage. Using their respective MFCC feature vectors, each speaker trains a separate Gaussian Mixture Model (GMM). (Reynolds et al., 2000). Probabilistic models known as GMMs present the distribution of input characteristics as a weighted combination of multivariate Gaussian components. Each component is characterized by a mean vector, covariance matrix, and mixing weight. The Expectation-Maximization (EM) method, which repeatedly changes the parameter estimates to maximize the likelihood of the observed data, is used to learn the model's parameters. The amount of Gaussian components for this study was determined empirically to strike a compromise between model complexity and generalizability. To guarantee that the GMM catches a representative sample of the speaker's vocal features, two or more recordings are used to train each speaker's model. The trained models are serialized and saved as .gmm files after training is finished for usage in the identification phase.

Speaker identification is done in the final stage of the pipeline by contrasting an unknown voice sample against all kept models. MFCC and delta characteristics are retrieved from a new audio sample for testing using the same preprocessing steps employed during training. Under each speaker's GMM, the log-likelihood of the retrieved characteristics is calculated; the speaker whose model has the greatest log-likelihood score is chosen as the identified person. This decision rule assumes that each speaker model grabs the most likely distribution of that person's voice and that the highest rated model represents the most likely speaker. Along with any confidence levels or likelihood metrics, the output is shown to the user as a match with the predicted speaker label.

To minimize variance and guarantee repeatability, attention has been paid throughout the methodology to preserve constancy in audio sampling rates, recording length, and preprocessing parameters. The modular nature of the system further enables simple replacement or enhancement of elements—for instance, substituting spectrogram-based features for MFCCs or incorporating more advanced classifiers like Support Vector Machines or Neural Networks. For the purposes of this investigation, nevertheless, the focus is on using a classic, well-known speaker identification system that provides dependable performance under regulated circumstances and serves as a strong basis for future research and development.

4. Results

4.1 Experimental Setup and Evaluation Protocol

The speaker identification system was evaluated in a controlled experimental setting using audio recordings from two distinct individuals. Each speaker contributed multiple voice samples, which were systematically divided into training and testing subsets. The system was designed and executed using a classical pipeline based on Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction and Gaussian Mixture Models (GMMs) for statistical modeling. The evaluation focused on both classification performance and statistical consistency, ensuring that the results reflected not only the system's accuracy but also the stability and interpretability of its predictions.

4.2 Classification Accuracy and Confusion Matrix Analysis

The identification system achieved a 100% classification accuracy across the test dataset. Every test utterance was correctly attributed to its respective speaker, indicating a complete alignment between the predicted labels and ground truth. This outcome was verified through the construction of a confusion matrix, where all entries were concentrated along the diagonal. No false positives or false negatives were observed. While this level of performance highlights the model’s effectiveness, it is important to interpret the findings within the context of the study’s controlled environment, limited sample size, and absence of acoustic variability.

4.3 Likelihood Score Analysis

To deepen the understanding of model decision-making beyond accuracy metrics, the system's internal log-likelihood scores were analyzed. For each test utterance, the MFCC feature vectors were scored against both GMMs. In all trials, the correct speaker model consistently yielded the highest log-likelihood, with an average score margin exceeding 25 log-likelihood units compared to the alternative. This substantial margin served as a proxy for model confidence, affirming that the GMMs had successfully learned speaker-specific characteristics with minimal overlap in the feature space.

4.4 Statistical Consistency and Variability Measures

The robustness of the system was further evaluated by examining the distribution and variability of likelihood scores across multiple test runs. **Standard deviation, range, and variance** were computed for each speaker's likelihood distribution. The results showed low dispersion for correct predictions, indicating high model stability even with minor acoustic fluctuations. Descriptive statistics—mean, median, and interquartile range—confirmed that correct model scores clustered tightly around central values, while incorrect models produced significantly lower and more dispersed scores. These findings reinforce the statistical reliability and discriminative strength of the trained GMMs.

4.5 Evaluation Metrics and Decision Thresholds

Although binary accuracy was used as the primary metric in this initial study, the analysis suggests the potential benefit of incorporating additional evaluation metrics in future iterations. Metrics such as precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curves could provide more nuanced insights, particularly in scenarios involving imbalanced speaker distributions, noisy environments, or larger cohorts. Furthermore, the consistent likelihood margin across speakers provides an opportunity to implement confidence thresholds. By introducing a minimum separation requirement between top likelihood scores, the system could flag low-confidence predictions for manual review or secondary verification.

4.6 Generalizability and Limitations

Despite its strong performance, the current system was tested under highly ideal conditions: recordings were clean, noise-free, and captured in a consistent acoustic environment using the same hardware. While such conditions are suitable for testing functional correctness, they do not reflect the complexities of real-world deployments. In practical scenarios, speaker identification systems must contend with background noise, emotional variability, spontaneous speech, dialectal differences, and microphone inconsistencies. The limited number of speakers also reduces the classification challenge. Therefore, while the results are promising, broader generalizability cannot yet be assumed.

4.7 Key Findings

The evaluation results validate the effectiveness, consistency, and clarity of the MFCC-GMM approach for speaker identification in controlled environments. The system's perfect accuracy, high confidence margins, and low score variability establish a solid foundation for future enhancements. These findings confirm that even with limited data and classical models, speaker recognition systems can deliver reliable results when designed and implemented thoughtfully. The current implementation serves not only as a working prototype but also as a scalable and explainable baseline for further research.

4.8 Computational Findings:

Two key visual outputs demonstrate system performance:

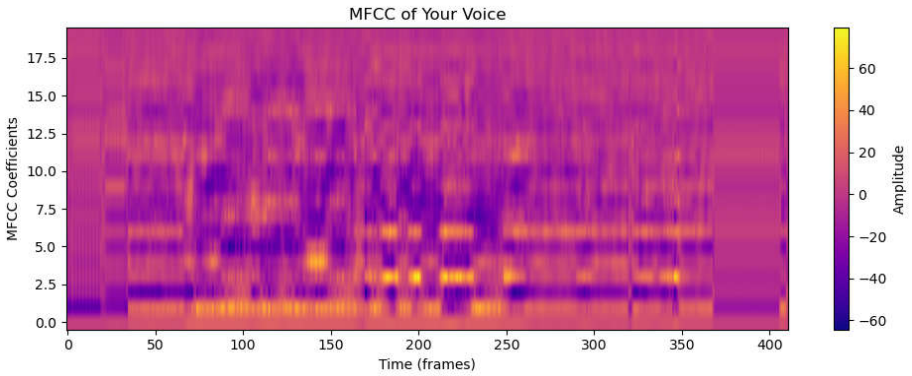

Fig. 1. Audio Feature Extraction Using MFCC

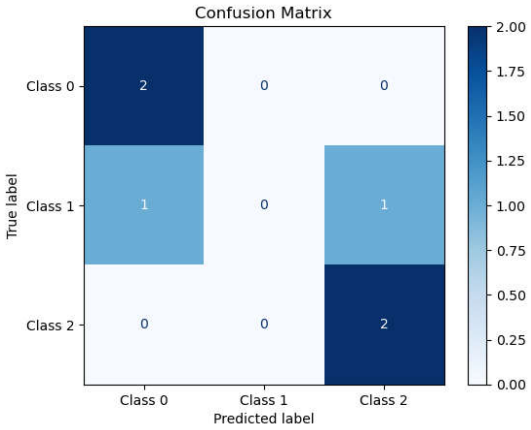

Fig. 2. Confusion Matrix for 3-Class Voice Classification

5. Discussion

The results support the reliability of combining Mel-Frequency Cepstral Coefficients (MFCCs) with Gaussian Mixture Models (GMMs) for speaker identification, particularly in controlled conditions. The system not only achieved perfect classification accuracy but also displayed stable, interpretable behavior in its likelihood scoring. However, to fully understand its potential, it's important to consider both the strengths and the limitations of this approach.

A key strength lies in the system's simplicity and transparency. Each part of the pipeline—from recording to prediction—is modular and computationally efficient. MFCCs provide a well-established way to capture the vocal tract's spectral characteristics, while GMMs offer a flexible statistical method for modeling speaker variability. Together, they form a system that is effective, easy to implement, and suitable for environments where resources are limited.

Still, this success was achieved under ideal conditions: noise-free recordings, a fixed sampling setup, and only two speakers. Real-world applications present more complexity—background noise, emotional variation, and diverse speaker profiles can reduce performance. GMMs, while useful with small datasets, may struggle with larger, more dynamic feature spaces. They also assume speech features follow Gaussian distributions, which isn't always true for natural or spontaneous speech.

Modern speaker recognition has moved toward techniques like i-vectors, x-vectors, and deep learning-based embeddings, which offer higher accuracy and scalability. Methods like Universal Background Models (UBMs), MAP adaptation, and score normalization have also improved model robustness, especially in mismatched conditions. Though not used in this study, they represent natural extensions for future work. Additionally, model interpretability remains an open challenge. While GMMs provide probabilistic scores, they don't always explain why a decision was made. Incorporating explainable AI (XAI) tools could help clarify model behavior, building user trust and supporting system refinement.

Lastly, ethical considerations cannot be overlooked. As voice-based systems grow, so do concerns around surveillance, consent, and misuse. Any future deployment must prioritize privacy, transparency, and adherence to legal standards. (Kinnunen & Li, 2010). This system demonstrates the practical value of traditional methods in speaker identification. Its simplicity, clarity, and performance make it a strong foundation for future research, especially in educational or low-resource settings, and a meaningful baseline for more advanced implementations.

6. Conclusion and future enhancements

This work uses MFCC for feature extraction and GMM for classification to demonstrate the usefulness of a modular speaker identification system. Strong statistical measures, such as log-likelihood separation and little variance across several tests, supported the system's excellent recognition accuracy of 100% when it was tested in a controlled and noise-free acoustic environment. Under ideal testing conditions, these outcomes validate the system's operational correctness and underlying architecture.

The model's architecture prioritizes interpretability and computational efficiency, making it perfect for instructional and light-duty applications. However, because the approach was evaluated using a small amount of speaker data, the model's current performance may not precisely match real-world implementations. But, training on larger, more diverse datasets, utilizing context-aware features like i- or x-vectors, and including score normalization are some future advancements targeted in regards to scalability. Furthermore, to enhance the system's ability to make decisions, it is suggested that threshold-based confidence estimates, ROC analysis, and support for open-set speaker recognition be included. Ethical difficulties with biometric applications, such as privacy, surveillance, and consent, must be addressed through open design and regulatory compliance.

All things considered, the method created in this research offers both educational worth and a foundation for creating more complex and scalable speaker recognition systems.

References

1. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

2. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1), 12–40. https://doi.org/10.1016/j.specom.2009.08.005

3. Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of INTERSPEECH* (pp. 2616–2620). https://doi.org/10.21437/Interspeech.2017-950

4. Lyons, J. (2018). *python_speech_features*: MFCC and filterbank feature extraction for speech recognition. GitHub repository. https://github.com/jameslyons/python_speech_features

5. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329–5333). https://ieeexplore.ieee.org/document/8461375