Voice Biometric Authentication System

Kushal Agrawal, Nachiketa Purohit

m23csa011@iitj.ac.in, m23csa016@iitj.ac.in

1. Introduction and Motivation

The quest for security in today's world, characterized by pervasive insecurity, stands as a paramount concern. Amidst this backdrop, voice biometrics emerges as a pivotal field, particularly in the realm of authentication. Leveraging the distinct attributes of human voice—comprising both physiological and behavioral characteristics—voice biometrics enables speaker recognition, facilitating the identification of individuals through their unique vocal signatures. Unlike traditional methods, voice biometrics transcends environmental variables or communication channels, offering a reliable and cost-effective means of authentication in the digital age. Its applicability spans various domains, including forensics, remote access control, web services, online communication, and customer relationship management, among others.

Forensic science, a cornerstone of crime investigation, has evolved to incorporate digital forensic techniques, aimed at gathering and analyzing digital evidence for legal purposes. Within this framework, forensic linguistics, also known as forensic phonetics, plays a crucial role in voice identification, utilizing voice acoustic properties to ascertain speaker identity. Forensic speaker recognition, a specialized application of speaker recognition, thus emerges as a pivotal tool in investigative endeavors.

Speaker recognition, encompassing both verification and identification, epitomizes the process of authenticating individuals based on their voice characteristics. This biometric identification technique, rooted in human traits, finds application across authentication, forensic analysis, and indexing scenarios. Authentication entails validating user identity for access control, while voice forensic endeavors to match voice samples for source attribution. Moreover, screening and indexing applications involve sifting through extensive voice databases to pinpoint specific speakers.

This paper aims to delve deeper into the realm of voice biometrics and speaker recognition. Following this introduction, the subsequent sections elucidate various biometric techniques and their classifications, expound upon the concept of speaker recognition technology and its operational phases, delve into prosodic features for speaker recognition, delineate the diverse applications of speaker recognition, and discuss the metrics employed to gauge the performance of biometric systems. Ultimately, the paper concludes by encapsulating key insights drawn from this exploration.

2. Problem Statement

In educational institutions, ensuring secure access to online learning platforms and academic resources is paramount to safeguarding sensitive student data and preserving the integrity of academic information. Traditional authentication methods, such as passwords or PINs, are susceptible to vulnerabilities such as theft, hacking, or unauthorized access. Consequently, there is a pressing need for robust authentication systems that offer heightened security measures while ensuring seamless access for authorized users.

The proposed project seeks to address this challenge by implementing a voice biometric authentication system tailored specifically for educational institutions. By leveraging the unique characteristics of individuals' voices, this system aims to provide a secure and user-friendly means for students to access online learning platforms and educational resources. Through voice biometric authentication, students can securely authenticate their identities, mitigating the risks associated with password-based authentication methods.

However, the implementation of voice biometric authentication in educational settings presents several challenges. One such challenge involves the development of accurate and reliable voice recognition algorithms capable of accurately verifying the identities of students across diverse speaking environments and communication channels. Additionally, concerns regarding privacy and data protection must be addressed to ensure compliance with regulatory frameworks and safeguard students' personal information.

Furthermore, the scalability and integration of voice biometric authentication systems within existing educational infrastructure require careful consideration to ensure seamless deployment and adoption. Factors such as system interoperability, user training, and resource allocation must be carefully managed to facilitate the successful implementation of voice biometric authentication in educational institutions.

Addressing these challenges is crucial to the successful implementation of voice biometric authentication systems in educational institutions, offering enhanced security measures while ensuring convenient access to online learning resources for students. By mitigating the risks associated with traditional authentication methods, voice biometric authentication has the potential to significantly improve the security posture of educational institutions and protect sensitive academic information from unauthorized access.

3. Literature Review

"Voice Biometric Identity Authentication System Based on Android Smart Phone" presents a comprehensive exploration of voice biometric authentication systems within the context of Android platforms, highlighting the increasing relevance of voice biometrics in the realm of mobile authentication due to the inherent vulnerabilities of traditional authentication methods. By utilizing voice databases such as TIMIT and XJTU

VOICE, the study underscores the pivotal role of robust voice data in the training and validation of voice recognition models, emphasizing the importance of data quality in achieving accurate and reliable authentication outcomes. Leveraging sophisticated signal processing techniques like Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and adopting a Mixed Gaussian function model for training voiceprint features, the system demonstrates a notable success rate ranging from 89 to 96 percent in authentication tasks, coupled with a real-time authentication time spanning from 210ms to 320ms. This performance showcases the system's efficacy in delivering high levels of accuracy and operational efficiency, thereby positioning voice biometrics as a promising and secure avenue for identity verification on Android smart phones, catering to the evolving needs of mobile security in contemporary digital landscapes. [1].

"Voice Biometric: A Technology for Voice Based Authentication" presents a technology utilizes both physiological and behavioral characteristics of the human voice for authentication purposes. Physiological characteristics include the physical aspects of the voice, while behavioral characteristics encompass the patterns and habits in an individual's speech. Voice biometric systems have the potential to authenticate individuals regardless of changes in the environment or communication channel, making them a cost-effective and easily accessible security solution in the digital era. The applications of speaker recognition systems, including voice biometric technology, are diverse and include areas such as forensics, remote access control security, web services, online calling, personalization of services, customer relationship management, voice-based banking, and surveillance/criminal investigation. Forensic linguistics, a method of voice identification based on voice acoustic qualities, plays a crucial role in forensic speaker recognition, which is an application of speaker recognition technology. The development and improvement of voice biometric systems are essential for enhancing accuracy and reliability, especially in critical fields such as voice indexing, medical records, online transactions, fraud prevention, and access control. Overall, the review of existing literature on voice biometric technology highlights its significance in providing secure authentication solutions and its potential applications across various industries and sectors.[2].

4. Proposed Solution

The process of speaker recognition, a subset of voice recognition within the broader field of machine learning and pattern recognition, entails distinguishing between different speakers and discerning the content of their speech. It is essential to differentiate between speaker recognition, which identifies "who is speaking," and speech recognition, which identifies "what is being said." This distinction underscores the importance of accurately extracting and analyzing speech features to develop effective speaker recognition systems.

At the core of speaker recognition lies the extraction of speech signals, a pivotal task that involves various techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and Prosodic analysis. These techniques enable the representation of distinct voice characteristics essential for identifying individual speakers. Once speech features are extracted, speaker models are constructed and stored within a voice database using diverse modeling techniques such as Gaussian Mixture Models (GMM)

The speaker recognition process encompasses two main phases: enrollment and verification. During enrollment, a speaker's voice is recorded, and specific features are extracted to create a unique voice print. In the verification phase, a voice sample is compared against the stored template or voice print to ascertain the speaker's identity. This process can be further categorized into speaker verification and speaker identification. Speaker verification involves comparing a voice sample against a claimed identity, akin to presenting an identity card for scrutiny. In contrast, speaker identification entails matching a voice sample against multiple templates to determine the best match, resembling the process of identifying a suspect from a database of criminals.

It is noteworthy that verification typically precedes identification in practice due to its faster processing time. Identification serves to narrow down potential matches, allowing for more efficient verification. For instance, when dealing with a suspected assailant, the identification phase aims to identify potential matches within the voice database, followed by verification to confirm the match's accuracy.

In summary, our approach encompasses understanding the intricacies of speaker recognition, from speech signal extraction to model creation and the enrollment-verification process. By delineating the distinct phases and techniques involved, we aim to elucidate the underlying mechanisms of speaker recognition systems and their practical applications in various domains.

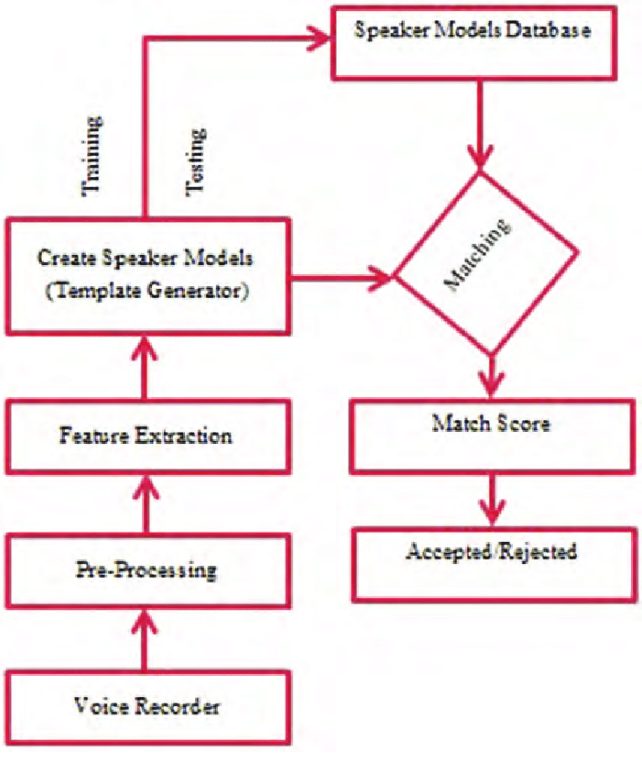

Figure 1: *Phases of automatic speaker recognition system*

5. Feature Extraction

Voice data encompasses a multitude of diverse feature quantities, each carrying its own corresponding physical significance. The selection and extraction of these feature vectors are pivotal in shaping the efficacy of voice authentication systems. An ideal feature set should exhibit several key characteristics:

1. **Discriminative Power**: The feature set should exhibit significant variance between different individuals while remaining relatively invariant for the same individual.

2. **Robustness**: Features should demonstrate resilience against noise and distortion commonly encountered in real-world environments.

3. **Natural Periodicity**: Features should capture the inherent periodicity present in voice signals, reflecting human speech patterns.

4. **Ease of Extraction**: Feature extraction should be computationally efficient and readily attainable from voice signals.

5. **Difficulty of Imitation**: Extracted features should be challenging to mimic, enhancing the security of voice authentication systems.

6. **Stability Over Time**: Features should exhibit consistency across different instances of the same speaker over time and remain unaffected by changes in the speaker's health.

Furthermore, to facilitate efficient processing and enhance discriminatory power, the number of extracted features should be minimized to focus solely on relevant information.

5.1. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) emerge as a prominent choice for voice feature extraction due to their ability to capture pertinent characteristics of human auditory perception. The MFCC coefficients are obtained in the Mel frequency domain, which closely mirrors the nonlinear frequency perception of the human ear. The transformation from the actual frequency f to the Mel frequency $\text{Mel}(f)$ is approximated by:

$$\text{Mel}(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right)$$

The feature extraction process based on MFCC entails several stages:

1. **Preemphasis**: The voice signal undergoes preemphasis through a high-pass filter to accentuate its high-frequency components, enhancing the spectral characteristics.
2. **Frame Division**: Due to the short-term stationary nature of voice signals, the signal is divided into shorter frames to capture local spectral characteristics. Each frame typically comprises a set of N sampling points, with an overlap region of M sampling points between adjacent frames.
3. **Windowing**: A window function is applied to each frame to emphasize the waveform near the sampling points while attenuating the rest. Common window functions include rectangular, Hamming, and Hanning windows.
4. **FFT Transformation**: Each windowed frame undergoes Fast Fourier Transform (FFT) to convert it from the time domain to the frequency domain, yielding its energy distribution in the frequency domain. The power spectrum is then calculated from the squared magnitudes of the FFT coefficients.
5. **MFCC Calculation**: Finally, MFCC features are computed from the power spectrum using a series of discrete cosine transformations, yielding a compact representation of the voice signal suitable for subsequent analysis.

5.2. Chroma Features

Chroma features are another essential component of voice feature extraction, particularly in music and tonal speech analysis. Chroma features represent the energy distribution of musical pitch classes within a frame of audio. This representation is invariant to changes in timbre and loudness, making it valuable for tasks such as music genre classification and chord recognition.

5.3. Root Mean Square (RMS) Features

Root Mean Square (RMS) features capture the average energy of a signal within each frame. These features provide insights into the overall amplitude characteristics of the voice signal, aiding in tasks such as speech detection and segmentation.

In addition to MFCC, Chroma, and RMS features, various other feature sets may also be explored to capture different aspects of the voice signal, depending on the specific requirements of the application.

This comprehensive feature extraction process enables the generation of discriminative feature parameters characterizing the unique voice signatures of individual users, forming the foundation for robust voice authentication systems.

5.4. Voice Modeling

Gaussian Mixture Models (GMM) will be employed to create statistical models that effectively capture variability in voice samples, enhancing the system's ability to authenticate users accurately.

5.5. Matching Algorithms

Advanced matching algorithms, including Maximum A Posteriori (MAP) estimation, will be implemented to improve the accuracy of matching voice features against enrolled voice models, ensuring reliable authentication outcomes.

6. Dataset

Svarah: Indian-English Speech Dataset

- Size: 9.6 hours of transcribed English audio
- Speakers: 117 diverse individuals
- Geographic Coverage: 65 districts across 19 Indian states
- Accent Diversity: Covers 19 of 22 constitutional languages (wide range of accents)
- Content: Includes both read speech and spontaneous conversations (provides diverse speaking styles)

7. Evaluation Metric

The Equal Error Rate (EER) is a commonly used metric for evaluating the performance of speaker verification systems. It represents the point where the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. In other words, it indicates the rate at which the system makes both types of errors (accepting an imposter and rejecting a genuine user) with the same frequency.Lower EER values correspond to better system performance.

For our voice biometric system focused on Indian English accents, EER will be a key metric to assess its ability to accurately identify and verify users while minimizing both false acceptances and false rejections [2].

8. References

[1] X. Zhang, Q. Xiong, Y. Dai, and X. Xu, "Voice biometric identity authentication system based on android smart phone," in *2018*

IEEE 4th International Conference on Computer and Communications (ICCC), 2018, pp. 1440–1444.

[2] N. Singh, A. Agrawal, and R. Khan, "Voice biometric: A technology for voice based authentication," *Advanced Science, Engineering and Medicine*, vol. 10, no. 7-8, pp. 754–759, 2018.