International Journal of Intelligent Communication and Computer Science
Vol. 2, No. 1, 2024, Pages 13-33.

# Real Voice Recognition and Authentication System: A Comprehensive Review

Review Paper

**Manish Tiwari[1] and Deepak Kumar Verma[2]**

[1]Research Scholar, Department of Computer Science and Engineering, University Institute of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur, U.P. INDIA
[2] Department of Computer Science and Engineering, University Institute of Engineering and Technology, Chhatrapati Shahu Ji Maharaj University, Kanpur, U.P. INDIA

Email: tiwarimanish1981@gmail.com

Received: 22 Jan 2024, Revised: 29 Mar 2023 Accepted: 12 Apr 2024

**Abstract:**

This comprehensive review paper explores the cutting-edge developments in the field of "Advancements in Real Voice Recognition and Authentication." As technology continues to evolve, the ability to accurately recognize and authenticate individuals through their real voice has become increasingly crucial for a variety of applications, including security systems, user identification, and access control. The review delves into recent advancements in speech recognition algorithms, acoustic-phonetic approaches, pattern recognition methodologies, template-based techniques, statistical models, and stochastic approaches applied to real voice recognition. The paper provides a thorough examination of the strengths and limitations of each approach, highlighting their contributions to the enhancement of authentication accuracy and system robustness. Additionally, it discusses emerging trends, challenges, and potential future directions in the dynamic landscape of real voice recognition and authentication technologies. This review aims to offer valuable insights for researchers, practitioners, and policymakers interested in the forefront of advancements in voice-based recognition and authentication systems.

**Keywords:** Voice Recognition, speaker recognition and real time

## 1. Introduction

Every individual possesses a distinctive and unique voice, shaped by a combination of physiological and behavioral components [1]. The physiological aspect is primarily associated with the structure of the vocal tract, while the behavioral component encompasses elements such as accent and tonal nuances. The amalgamation of these factors results in a singular voice profile that can be employed for precise identification [2]. To harness the inherent individuality of voices, advanced biometric voice recognition systems have been developed for user authentication, relying solely on the analysis of voice samples [3]. A focal point in many voice recognition systems is the vocal tract, which exhibits distinctive characteristics specific to each person, forming the basis for accurate identification. Illustrated in Figure 1, is a representation of a voice verification system, highlighting the growing significance of this technology. Notably, the advantages of utilizing such systems are noteworthy. Firstly, the cost-effectiveness of this technology makes it an attractive option. Additionally, the ubiquity of telephones allows for the seamless integration of voice recognition techniques. However, it is essential to acknowledge the limitations and drawbacks of biometric voice recognition. One notable disadvantage is the potential for voice mimicry, introducing vulnerability in the system [4]. Moreover, individuals with a sore throat may encounter challenges in accessing their information, posing a practical hurdle for authorized users [5]. The natural aging process also contributes to changes in voice over time, further complicating the accuracy of identification [5]. Furthermore, the presence of background or foreground noise in the environment can pose difficulties in accurately recognizing voices, adding another layer of complexity to the technology [6]. Despite these challenges, it is crucial to recognize that biometrics, including voice recognition, does not guarantee absolute security. However, it does offer a convenient and reliable method for user identification and authentication. As technology continues to evolve, addressing these challenges will be imperative to enhance the effectiveness and reliability of biometric voice recognition systems in various applications.

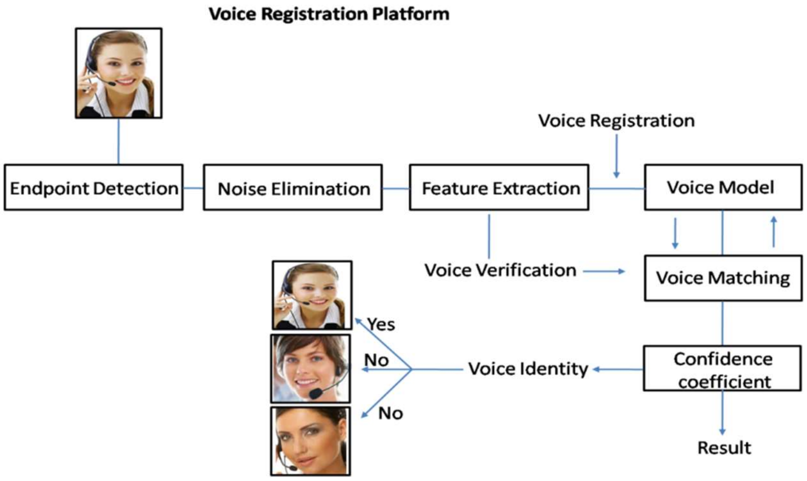


**Figure 1: Voice verification system**

### Speaker Identification "Si"

Speaker identification is the process of discerning and attributing a person among a multitude of speakers by evaluating and comparing their vocal features against established reference levels [7]. In the depicted Automatic Speech Recognition (ASR) system, as illustrated in Figure 2, an input voice is introduced to the system and subjected to a comparison with pre-recorded voices of known speakers stored in the system's database. The system conducts a meticulous assessment, considering not only the specific words uttered but also the arrangement of these words, in order to establish the identity of the speaker. This involves a thorough comparison with the characteristic references associated with each known speaker in the system. In the intricate process of speaker identification, the ASR system essentially acts as a discerning arbiter, analyzing the acoustic features and speech patterns of the input voice against the stored references. The system's output, in turn, provides the identity of the speaker whose reference values closely align with the uttered sequence of words. By leveraging this approach, speaker identification not only enhances the security and authentication aspects of voice-based systems but also finds applications in diverse fields such as telecommunications, security access, and personalized user interfaces.

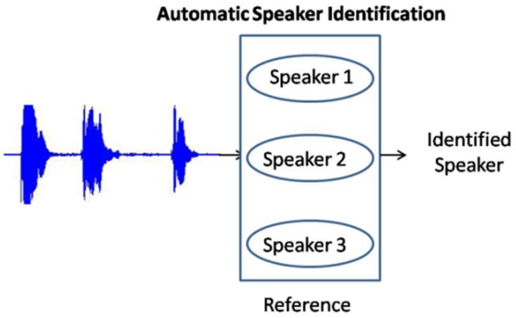


**Figure 2: Schematic of Speaker identification system**

**Speaker Verification "Sv"**

The authentication of a speaker is an imperative step preceding user access, necessitating a meticulous examination of their vocal message in comparison to the established acoustic reference associated with the claimed identity (as depicted in Figure 3). This authentication process entails the calculation of a similarity measure between the vocal message of the incoming user and the reference values attributed to the asserted speaker [8]. The computed similarity measure is then juxtaposed with a predetermined threshold value.

In this discerning procedure, if the similarity measure surpasses the predefined threshold, the speaker is duly accepted, affirming the claimed identity. Conversely, if the similarity measure falls below the stipulated threshold, the system rejects the speaker, deeming them an impostor. This intricate yet crucial verification mechanism ensures that only authorized speakers gain access, bolstering the security and integrity of speaker authentication systems. Such procedures find extensive application in diverse domains where user verification based on vocal characteristics is pivotal, including secure access control systems, voice-enabled devices, and biometric authentication protocols.

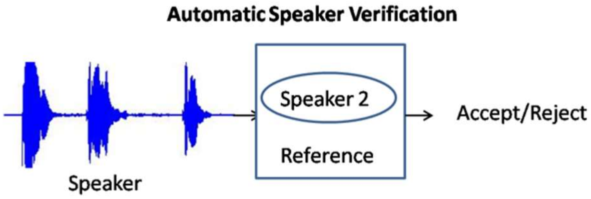


**Figure 3: Schematic of Speaker verification system**

## 2. Voice Recognition

Speech stands out as the most natural, efficient, and fundamental means of communication among humans, making it a logical progression that innovation would lead to the development of speech recognition [9]. Speech recognition can be defined as the process of converting speech signals into a sequence of words through algorithms implemented as computer programs. In contemporary times, with advancements in statistical speech modeling, the demand for automatic speech recognition has been steadily rising [10]. This is particularly evident in applications such as automatic call processing, where a human-machine interface is crucial. In the early stages of development, even basic tasks like digitizing (sampling) voice posed significant challenges. The breakthrough came in the 1980s when the first systems capable of deciphering speech emerged, albeit with limited scope and capabilities. Speech recognition involves generating a word sequence that accurately corresponds to the given speech signal. Prominent applications include auto-attendants, natural language understanding, virtual reality, multimedia searches, travel information and reservations, translators, and numerous others [12]. Notably, contemporary technologies like Google's speech-based search and Microsoft Windows 8's speech recognition system for user login on PCs and laptops exemplify the integration of speech recognition into everyday applications [13].

## 3. Speech Recognition System

The schematic diagram illustrating the speech recognition system is presented in Figure 4. Within this diagram, the distinct stages of the system, such as pre-processing, feature extraction, and modeling, are depicted during the training phases. In the testing phase, the diagram highlights the steps of pre-processing, feature extraction, and pattern matching. Further elaboration on each of these processes is provided in the subsequent sub-sections for a comprehensive understanding of the speech recognition system's functionality.

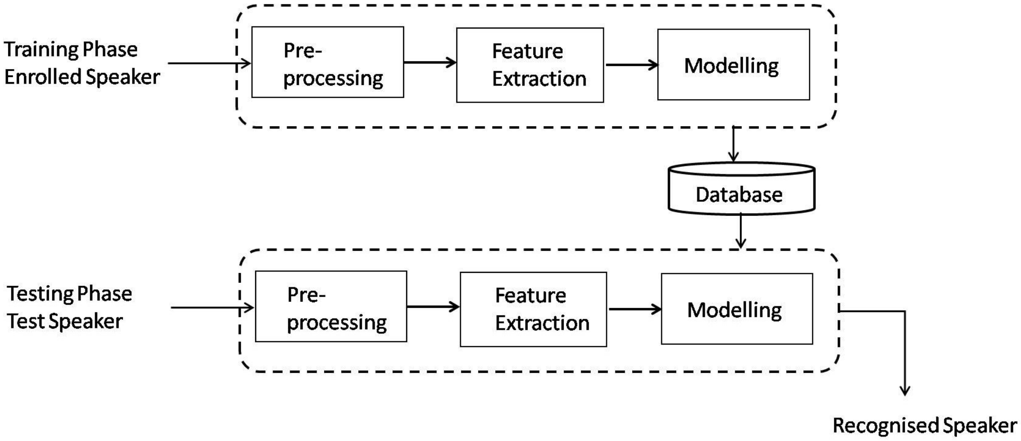


**Figure 4: Schematic of Speaker recognition system**

### 3.1 Pre-processing Steps:

**(a) Sampling**

Sampling serves as the foundational step in the pre-processing pipeline for speech recognition [14]. In this crucial stage, the continuous analog signals, which constitute the raw audio input, undergo a transformation into a discrete digital form. This transformation is imperative for subsequent digital analyses and computations. Sampling involves capturing the amplitude of the audio signal at predetermined regular intervals, transforming the continuous waveform into a sequence of discrete data points [15]. Each data point corresponds to a specific instance in time, providing a quantized representation of the audio signal. This quantization not only facilitates the handling of audio data in digital systems but also sets the stage for a range of subsequent signal processing techniques [16]. By converting the analog signal into a digital format through sampling, the speech recognition system gains the ability to analyze and interpret the audio input with precision, laying the groundwork for further pre-processing steps and subsequent recognition tasks.

**(b) Normalization**

Normalization is crucial to maintaining consistent volume levels across different recordings. In this step, the amplitude of the audio signal is adjusted to a standardized level [17]. This ensures that variations in recording volume do not adversely affect the performance of the speech recognition system, allowing for more reliable analysis.

**(c) Filtering**

In the pre-processing phase of speech recognition, filtering emerges as a pivotal step aimed at elevating the quality of the audio data. The primary objective is twofold: to eliminate unwanted noise and accentuate pertinent frequency components inherent in the speech signal [18]. The application of filtering techniques is instrumental in refining the overall signal clarity, paving the way for more accurate subsequent analyses. High-pass and low-pass filters, recognized for their efficacy, are commonly employed in this context. High-pass filters selectively permit the transmission of frequencies above a specified threshold, effectively attenuating or eliminating lower frequencies associated with background noise. On the other hand, low-pass filters allow frequencies below a designated threshold to pass through, retaining the essential components of the speech signal while mitigating higher-frequency interference. This strategic combination of high-pass and low-pass filtering not only contributes to noise reduction but also ensures that the speech frequencies, critical for accurate recognition, are preserved and emphasized [19]. By leveraging these filtering mechanisms, the pre-processing pipeline optimally prepares the audio data for subsequent stages in the speech recognition system, ultimately enhancing its ability to discern and interpret spoken content with precision.

**(d) Pre-emphasis**

In the spectrum of pre-processing techniques for speech recognition, pre-emphasis assumes a critical role by addressing a specific concern associated with audio recording [20] (Figure 5). The process of pre-emphasis is designed to compensate for the loss of high-frequency energy that may occur during the recording phase. As an audio signal is captured and transmitted, certain high-frequency components tend to experience attenuation, resulting in an uneven distribution across the frequency spectrum. Pre-emphasis counteracts this effect by selectively boosting the amplitudes of higher frequencies within the audio signal. This deliberate emphasis on higher-frequency components serves to balance the overall spectral characteristics of the signal. The significance of pre-emphasis becomes particularly pronounced in the context of speech recognition, where the accurate detection of speech patterns relies on a well-balanced representation of the frequency spectrum [21]. By enhancing the presence of high-frequency elements, pre-emphasis contributes to the overall improvement of signal quality. The net effect is an audio signal that not only retains its fidelity but also aligns more effectively with the characteristics that are crucial for successful speech pattern recognition. Thus, pre-emphasis stands as a strategic maneuver in the pre-processing workflow, fortifying the system's capability to detect and interpret speech nuances with heightened accuracy.

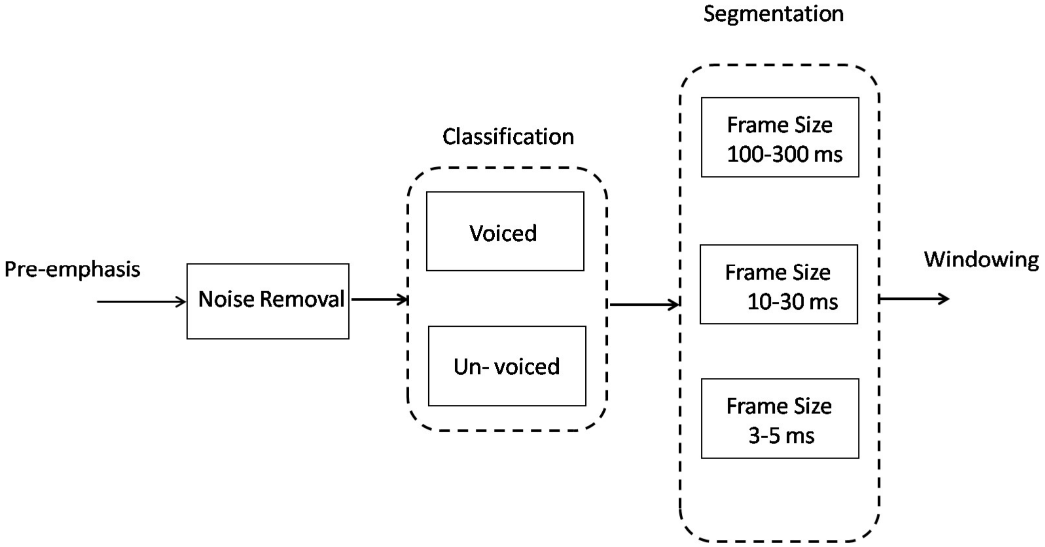


**Figure 5: Schematic of Speaker recognition processes**

### (e) Frame Blocking

Frame blocking stands as a pivotal stage in the pre-processing journey of speech recognition, introducing a strategic segmentation of the continuous audio signal. In this phase, the seamless waveform is partitioned into discrete, short frames, each representing a specific temporal segment for subsequent analysis [22]. These frames are meticulously crafted to have durations typically ranging between 20 to 30 milliseconds, a duration carefully chosen to capture meaningful phonetic information while maintaining temporal resolution. To ensure coherence and continuity in the representation of the audio signal, a degree of overlap is introduced between successive frames. This overlap allows for a smoother transition and integration of information across adjacent frames, preventing the loss of critical details at frame boundaries [23]. Frame blocking, therefore, serves a dual purpose: it breaks down the continuous signal into manageable temporal units, and the introduced overlap ensures a holistic representation of the audio signal. The system, in this segmented form, can then focus on these smaller units, facilitating a more granular and detailed analysis of the speech signal. This segmentation strategy aligns with the intricate nature of speech patterns, allowing the recognition system to delve into the nuances of each frame for a more robust understanding of the spoken content.

### (f) Windowing

In the pre-processing continuum of speech recognition, the application of windowing emerges as a critical refinement step, strategically employed to minimize spectral leakage during the processing of frames [24]. After the segmentation of the continuous audio signal into discrete frames, the inherent challenge lies in potential distortion at the frame edges due to abrupt transitions. Windowing addresses this challenge by introducing a window function, such as Hamming or Hanning, as a multiplier applied to each frame [25] (Figure 5). This process essentially shapes the amplitude of the frame, gradually tapering it towards the edges. The gradual tapering mitigates abrupt changes and reduces the spectral leakage that may occur during the analysis of individual frames. By modulating the amplitude in this manner, windowing ensures a smoother transition between frames, contributing to a more coherent and accurate representation of the audio signal. The choice of window function, whether it be the Hann window, Hamming window, or others, is guided by specific characteristics suited to the nuances of the speech signal and the nature of the subsequent analyses. Ultimately, the windowing process enhances the precision of the subsequent analytical steps, allowing for a more reliable and artifact-free interpretation of the speech signal in the realm of speech recognition.

### (g) Fast Fourier Transform (FFT)

In the progression of speech recognition pre-processing, the application of Fast Fourier Transform (FFT) marks a pivotal transition, transforming the audio signal from its native time domain into the frequency domain [26]. This transformation is instrumental in unlocking a detailed analysis of the frequency components embedded within the signal. The FFT algorithm dissects the complex waveform of the audio signal into its constituent frequencies, revealing the spectrum of the signal in a comprehensive manner [27]. Each frequency component is identified and quantified, laying the foundation for a more nuanced understanding of the spectral characteristics of the speech signal. The conversion from time domain to frequency domain is particularly crucial in speech recognition, as it enables the system to discern and interpret the specific frequencies associated with different phonetic elements. The FFT process essentially unveils the intricate frequency composition of the signal, providing a detailed representation that becomes the basis for subsequent analyses [28]. In essence, FFT is a transformative step that enhances the system's ability to delve into the rich frequency landscape of the speech signal, contributing to the accuracy and efficacy of the overall speech recognition process.

## 4. Feature Extraction Technique

In a categorization problem, particularly in the context of speaker verification and identification systems, speech feature extraction plays a crucial role [29]. The primary objective of this process is to reduce the dimensionality of the input vector while preserving the essential signal discriminating power. The rationale behind this lies in the inherent relationship between the dimensionality of the input and the number of training and test vectors required for effective classification. The basic formation of speaker verification and identification systems reveals that as the dimensionality of the input increases, the demand for a larger number of training and test vectors grows accordingly [30]. This phenomenon poses practical challenges, such as increased computational complexity and the need for more extensive datasets. Feature extraction addresses this challenge by condensing the relevant information within the speech signal, allowing for a more efficient and focused representation of the data. By extracting pertinent features from the speech signal, the system can retain the discriminative aspects essential for accurate classification while discarding redundant or less informative components. This not only facilitates more streamlined processing but also enhances the system's ability to generalize patterns from the training data to effectively classify new and unseen test data.

**Linear Predictive Coding (LPC)** stands as a fundamental technique in the realm of speech processing, dedicated to representing the spectral envelope of a given signal [31]. By modeling the vocal tract as a linear filter, LPC endeavors to estimate coefficients that effectively reconstruct the original signal [32]. This methodology finds extensive applications in diverse fields such as speech coding, speech recognition, and voice analysis. One of its notable advantages lies in its ability to compress speech signals efficiently while retaining vital information about the distinctive characteristics of the speaker [33].

**Linear Predictive Cepstral Coefficients (LPCC)** build upon the principles of LPC by introducing cepstral analysis into the feature extraction process [34]. This extension involves combining linear predictive coding with cepstral domain information, thereby enabling the capture of both short-term and long-term features inherent in speech signals [35]. LPCC has proven itself invaluable in various applications, including speaker recognition and speech synthesis, where a more detailed representation of spectral characteristics is imperative [36].

**Mel-Frequency Cepstral Coefficients (MFCC)** emerge as a highly favored technique for feature extraction in audio signals, particularly in the domains of speech and speaker recognition [37,38]. The MFCC process involves transforming the audio signal into the frequency domain through methods such as the Fourier transform, mapping the resulting spectrum onto the mel scale to simulate the non-linear response of the human auditory system, and finally computing cepstral coefficients [39] (Figure 6). Widely employed in automatic speech recognition and speaker identification, MFCC excels in leveraging the mel-scale to highlight frequency components that hold perceptual significance. Its applications extend to music information retrieval and various contexts where a robust representation of acoustic features is paramount [40].

**Table 1: Comparison of Feature extraction methods**

| Feature | LPC | LPCC | MFCC |
|---|---|---|---|
| Nature of Technique | Spectral envelope representation | LPC combined with cepstral analysis | Frequency domain representation on a mel scale |
| Modeling Approach | Linear filter model | Linear filter with cepstral analysis | Non-linear mapping onto mel scale |
| Key Strengths | Efficient signal compression, preserves speaker characteristics | Captures both short-term and long-term features, detailed spectral information | Perceptually relevant representation, effective in speech and speaker recognition |
| Applications | Speech coding, speech recognition, voice analysis | Speaker recognition, speech synthesis, detailed spectral analysis | Automatic speech recognition, speaker identification, music information retrieval |
| Operational Steps | Estimate coefficients for linear filter model | Incorporate cepstral analysis into LPC, capture both temporal features | Transform signal to frequency domain, map onto mel scale, compute cepstral coefficients |
| Perceptual Considerations | Focus on spectral envelope | Captures both spectral and temporal aspects | Emphasizes perceptually relevant frequencies |
| Usability | Broad range of speech applications | Speaker-specific applications, detailed analysis | Wide-ranging applications in speech and audio processing |

In essence, feature extraction in the context of speaker verification and identification is a strategic approach to strike a balance between maintaining the discriminatory power of the signal and managing the computational demands associated with higher-dimensional input vectors. It serves as a critical step in optimizing the efficiency and effectiveness of the classification process, enabling the development of more robust and practical speech-based recognition systems.

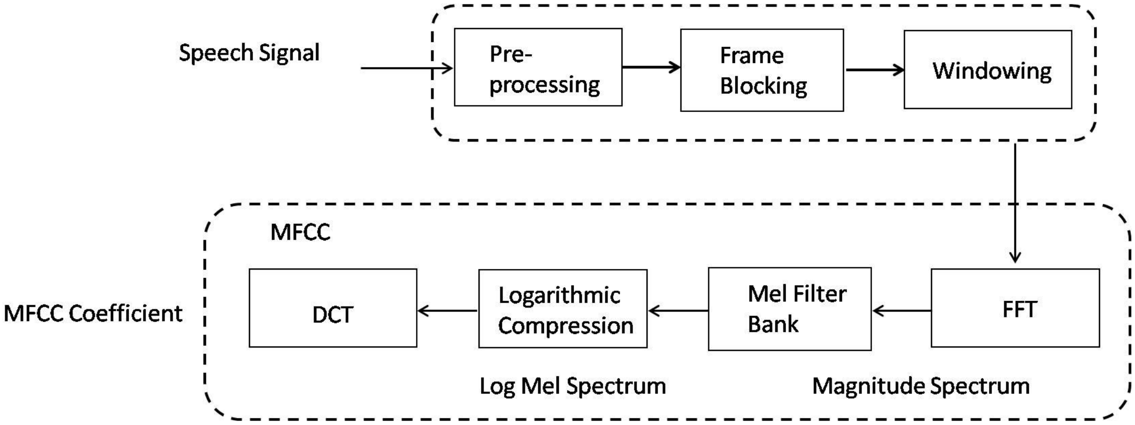


**Figure 6: Schematic of feature extraction process (MFCC)**

## 5. Modeling Technique

The technique mentioned involves the development of speaker models based on speaker-specific feature vectors and is broadly categorized into two classes: (a) speaker recognition and (b) speaker identification. Speaker identification, grounded in individual information embedded in speech signals, automatically discerns the identity of the speaker. Speaker recognition is further divided into two sections: (i) Speaker-dependent and (ii) Speaker-independent. In speaker-independent mode, the computer is tasked with disregarding speaker-specific qualities in the speech signal, extracting the intended message. Conversely, in speaker-dependent mode, the machine extracts speaker characteristics from the acoustic signal. The primary objective of speaker identification is to compare a speech signal from an unknown speaker to a database of known speakers, with the system capable of recognizing the speaker, having been trained on various speakers. Speaker recognition methods are further classified into text-dependent and text-independent approaches. In the text-dependent method, the speaker articulates specific keywords or sentences during both training and recognition trials. On the other hand, the text-independent method does not rely on specific texts being spoken by the speaker, allowing for more flexibility in recognizing speakers across diverse utterances.

Various modeling techniques can be employed in the speech recognition process:

### 5.1 Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (GMMs) serve as a prevalent choice for speaker modeling in the field of speech recognition [41]. They excel in representing the probability distribution of features by combining Gaussian distributions, allowing for the effective modeling of complex data inherent in speech signals [42]. This unique capability enables GMMs to capture the nuanced variations in acoustic features, such as pitch, intonation, and phonetic nuances, specific to different speakers [43]. The amalgamation of multiple Gaussian distributions within the model provides adaptability, making it well-suited for tasks like speaker identification and biometric systems. GMMs stand out as a powerful tool, offering a versatile and sophisticated approach to speaker modeling that contributes significantly to the advancement of speaker recognition applications [44].

### 5.2 Hidden Markov Models (HMMs):

Hidden Markov Models (HMMs) find application in speech signal processing, specifically to model temporal dependencies [45]. Their particular utility lies in effectively capturing the sequential characteristics inherent in speech features. As dynamic processes, speech signals unfold over time, and HMMs prove invaluable in encapsulating the evolving nature of these acoustic features [46]. By incorporating hidden states that evolve over time, HMMs are adept at modeling the transitions and dependencies between different states, allowing them to mirror the temporal progression of speech patterns. This sequential modeling capability is fundamental for tasks such as speech recognition, where understanding the temporal evolution of phonetic elements and linguistic structures is crucial for accurate and context-aware processing. HMMs, therefore, serve as a robust framework for encoding the dynamic nature of speech signals and extracting meaningful information from their temporal dependencies [47].

### 5.3 Neural Networks:

In the landscape of speaker recognition, deep learning techniques, notably neural networks, have emerged as prominent tools, revolutionizing the field [48]. Among these, Convolutional Neural Networks (CNNs) [49] and Recurrent Neural Networks (RNNs) [50] stand out as exemplary architectures that harness the power of deep learning to extract pertinent features.

Convolutional Neural Networks (CNNs) are particularly adept at processing grid-like data, making them well-suited for tasks involving spatial relationships. In speaker recognition, CNNs excel in capturing distinctive patterns within spectrograms or other two-dimensional representations of speech signals [49]. The convolutional layers in CNNs efficiently recognize local patterns, enabling them to discern key features indicative of individual speakers.

On the other hand, Recurrent Neural Networks (RNNs) are tailored to handle sequential data by incorporating memory elements. This makes RNNs highly effective in capturing the temporal dependencies present in speech signals, where the order of phonetic elements and the evolution of acoustic features over time are crucial for accurate recognition [50]. RNNs' ability to retain information from previous time steps allows them to model the sequential nature of spoken language.

These deep learning architectures have brought about a paradigm shift in speaker recognition, allowing systems to automatically learn and extract intricate features directly from the raw speech data [51]. The hierarchical and adaptive nature of neural networks enables them to discern complex patterns, contributing to heightened accuracy and robustness in speaker identification tasks. As deep learning continues to advance, these techniques pave the way for more sophisticated and context-aware speaker recognition systems, showcasing the transformative impact of neural networks in the realm of acoustic pattern recognition.

### 5.4 Vector Quantization (VQ):

Vector Quantization (VQ) is a technique employed in speech processing where speech feature vectors are mapped to a predefined set of representative codewords [52]. This mapping process enables the efficient and compact representation of speech signals, a crucial aspect in various applications such as speech recognition and signal compression.

In the context of speech processing, feature vectors are extracted from speech signals to capture essential information about the acoustic characteristics. VQ operates by associating these feature vectors with a set of codewords, each representing a prototypical pattern or cluster within the feature space [53]. The mapping is achieved by assigning each feature vector to the closest codeword, effectively quantizing the continuous feature space into a discrete set of representative points.

The advantages of VQ lie in its ability to reduce the dimensionality of the feature vectors, leading to a more streamlined and memory-efficient representation [54]. By grouping similar feature vectors into clusters represented by codewords, VQ condenses the information while preserving the essential characteristics of the speech signal. This not only aids in efficient storage but also facilitates quicker processing in applications such as real-time speech recognition.

Furthermore, VQ is instrumental in minimizing the impact of noise and variations in speech signals [55]. The discrete representation provided by codewords makes the system more robust to variations in pronunciation, accent, or environmental conditions, enhancing the overall reliability of speech processing systems.

### 5.5 Support Vector Machines (SVMs):

Support Vector Machines (SVMs) find valuable application in speaker recognition, specifically in tasks related to classification. These machines are instrumental in the delineation and separation of distinct speaker classes, thereby enhancing the discriminative power of speaker recognition systems [56]. The primary goal of SVMs in this context is to create a decision boundary that effectively distinguishes between different speaker categories based on the extracted features. In speaker recognition, SVMs contribute significantly to the modeling process by leveraging their capability to identify optimal hyperplanes in feature space, maximizing the margin between different speaker classes [57].

**Table 2: Comparison of Modeling Techniques**

| Method | Nature of Model | Strengths | Weaknesses |
|---|---|---|---|
| Gaussian Mixture Models (GMMs) | Probabilistic | Effective for modeling complex distributions | May struggle with capturing temporal dependencies |
| Hidden Markov Models (HMMs) | Temporal Modeling | Excellent for modeling temporal dependencies | May not capture long-term dependencies well |
| Neural Networks | Non-linear Modeling | Adaptability to complex patterns and high-dimensional data | Require substantial amounts of data for training |
| Vector Quantization (VQ) | Quantization | Compact representation of speech features | Sensitivity to variations in feature space |
| Support Vector Machines (SVMs) | Discriminative | Effective in high-dimensional feature spaces | Can be computationally intensive for large datasets |

This not only aids in accurate classification but also enhances the generalization ability of the model, allowing it to perform effectively on unseen data. The pivotal role of SVMs lies in their ability to handle complex, high-dimensional feature spaces efficiently [58]. By transforming the original feature space into a higher-dimensional space, SVMs enable the creation of decision boundaries that can effectively separate speakers based on their unique acoustic characteristics. This aspect is particularly crucial for speaker recognition systems aiming for robust and accurate performance across diverse datasets. Moreover, SVMs contribute to the robustness of speaker recognition systems against variations in speech patterns, accents, and environmental conditions. The discriminative power of SVMs assists in mitigating the impact of confounding factors, thereby improving the overall reliability of the system [59].

## 6. ASR System Classification

Speech recognition, a specialized domain within pattern recognition, is a multifaceted process characterized by two essential stages: Training and Testing. These phases operate within the framework of supervised pattern recognition, where the overarching goal is to accurately classify and identify spoken language patterns. Throughout both stages, a fundamental and shared procedure involves the extraction of features deemed pertinent for effective classification.

### Training Phase:

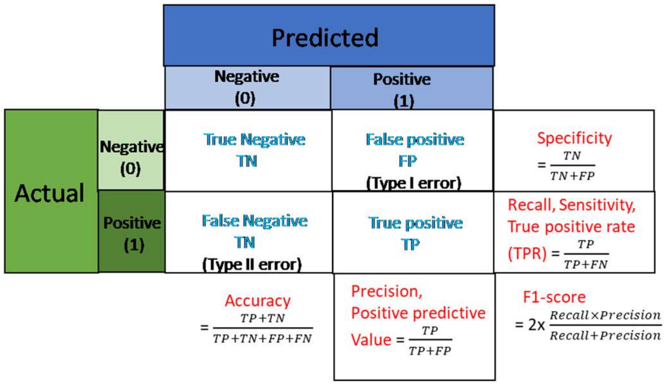
During the Training phase, the classification model parameters are assessed. This is achieved by utilizing an extensive set of class examples, known as Training Data. In this phase, the system learns to recognize patterns and variations within the provided data. The goal is to enable the model to accurately classify and differentiate between various classes based on the extracted features. The model's capacity to generalize and discern patterns in unfamiliar data improves proportionally with the diversity and representativeness of the training dataset.

### Testing Phase:

In the Testing phase, the learned model is put to the test with new, unseen data, referred to as test speech data. The process involves matching the features of the test pattern with the trained model for each class. The model essentially acts as a reference guide, and the test pattern is compared against it to identify the class it best aligns with. This matching process is crucial for determining the class to which the test pattern belongs. The decision regarding the class assignment is made based on the extent of similarity or matching between the features of the test pattern and the learned model for each class. In essence, the system determines which class's model provides the best match for the given test pattern. This matching process is fundamental to the accuracy and effectiveness of the speech recognition system.

### 6.1 Performance Measures

A confusion matrix is a table that is used to evaluate the performance of a classification algorithm. It provides a summary of the performance of a machine learning model by presenting the counts of true positive, true negative, false positive, and false negative values (Figure 7).



| | | Predicted | | |
|---|---|---|---|---|
| | | Negative<br>(0) | Positive<br>(1) | |
| Actual | Negative<br>(0) | True Negative<br>TN | False positive<br>FP<br>(Type I error) | Specificity<br>$= \frac{TN}{TN+FP}$ |
| | Positive<br>(1) | False Negative<br>TN<br>(Type II error) | True positive<br>TP | Recall, Sensitivity,<br>True positive rate<br>(TPR) = $\frac{TP}{TP+FN}$ |
| | | Accuracy<br>$= \frac{TP+TN}{TP+TN+FP+FN}$ | Precision,<br>Positive predictive<br>Value = $\frac{TP}{TP+FP}$ | F1-score<br>= 2x $\frac{Recall \times Precision}{Recall+Precision}$ |

**Figure 7: Schematic of confusion matrix**

**Accuracy:** Accuracy is a measure of the overall correctness of a classification model. It is calculated as the ratio of correctly predicted instances to the total instances.

$$\text{Accuray}=\frac{TP+TN}{TP+TN+FP+FN}$$

**Equal Error Rate (EER):** The Equal Error Rate is a point on the Receiver Operating Characteristic (ROC) curve where the false positive rate (FPR) equals the false negative rate (FNR). Mathematically, it is the point where the error rates for false positives and false negatives are equal.

The False Positive Rate (FPR) is given by:

$$\text{FPR}=\frac{FP}{FP+TN}$$

The False Negative Rate (FNR) is given by:

$$\text{FNR}=\frac{FN}{TP+FN}$$

The Equal Error Rate is the point where FPR=FNR

These formulas are used to quantify the performance of classification models. It's important to note that while accuracy provides an overall measure of correctness, the Equal Error Rate takes into account the trade-off between false positives and false negatives, offering a balanced evaluation of a classifier's performance.
ROC stands for Receiver Operating Characteristic. It is a graphical representation of the performance of a binary classification model as its discrimination threshold is varied. The ROC curve is created by plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings. In an ROC curve, each point represents a different discrimination threshold, and the curve illustrates the trade-off between sensitivity and specificity. A diagonal line (the line of no-discrimination) represents a model that makes random guesses, while a curve above the diagonal indicates a better-than-random performance. The Area Under the ROC Curve (AUC-ROC) is a summary measure often used to quantify the overall performance of a classification model. A higher AUC-ROC value indicates better discrimination ability.

## 7. Notable Works

**Jawarkar and Basu (2011)** delved into the utilization of a fuzzy min-max neural network for the purpose of speaker identification. The primary focus of their research was to investigate the efficacy of this neural network in capturing and discerning distinctive characteristics inherent to individual speakers. By employing the fuzzy min-max neural network, the researchers sought to enhance the overall accuracy and robustness of speaker identification processes. This involved a comprehensive exploration of how the network could effectively analyze and interpret the unique features within audio signals, ultimately contributing to the improvement of speaker identification methodologies. The study aimed to provide valuable insights into the potential applications and advantages of fuzzy min-max neural networks in the domain of speaker recognition, paving the way for advancements in the precision and reliability of automated speaker identification systems.

**Nakagawa et.al (2011)** embarked on a comprehensive exploration of speaker identification and verification by integrating Mel-frequency cepstral coefficients (MFCC) with phase information. The central aim of their study was to synergistically leverage both spectral and temporal features within audio signals to enhance the overall reliability and precision of speaker recognition systems. The researchers focused on the integration of MFCC, a widely used spectral feature extraction technique, with phase information, which captures temporal characteristics of the audio signal. This innovative approach aimed to capitalize on the complementary nature of spectral and temporal features, recognizing that they together provide a more holistic representation of speaker-specific characteristics. By combining these two types of features, Nakagawa et al. sought to address the inherent limitations of relying solely on spectral information for speaker recognition. The inclusion of phase information introduced a temporal dimension, enabling the model to capture nuances in speech patterns that might not be adequately represented by spectral features alone.

**Krishnamoorthy et. al (2011)** undertook a significant exploration in the field of speaker recognition by addressing the challenge posed by limited data conditions. In their study, the researchers proposed a novel approach to mitigate the effects of data scarcity by introducing controlled noise into the audio signals used for speaker identification [62]. The primary objective of their research was to investigate the impact of noise addition on the performance of speaker identification systems. Limited data scenarios are common in real-world applications, and they can pose challenges for accurate and reliable speaker recognition. By introducing noise deliberately, Krishnamoorthy et al. aimed to simulate conditions that more closely resemble the complexities of the natural acoustic environment. The study sought to understand how the presence of noise influences the robustness and adaptability of speaker identification systems. The researchers explored whether the introduction of noise could enhance the system's ability to generalize and accurately identify speakers even when confronted with variations in the acoustic environment.

**Tolba (2011**) made a significant contribution by proposing a high-performance, text-independent approach specifically designed for Arabic speakers. The foundation of this novel method was built upon the utilization of Continuous Hidden Markov Models (CHMM) [63]. The research was driven by a core objective: to advance speaker identification methodologies with a tailored focus on the linguistic characteristics inherent to Arabic speakers. The adoption of a CHMM-based approach in the proposed system reflected a sophisticated modeling technique. Hidden Markov Models (HMMs) are widely used in speech and speaker recognition due to their ability to represent sequential data. The continuous variant, CHMM, extends this capability to handle continuous observation streams, making it particularly suitable for the dynamic and diverse nature of speech signals. The significance of this research lay in its commitment to addressing the unique linguistic traits and challenges associated with Arabic speakers. Arabic, characterized by its distinctive phonetic features and phonological nuances, presents specific complexities that require specialized treatment in speaker identification systems. By tailoring the approach to the intricacies of Arabic speech patterns, Tolba aimed to enhance the overall accuracy and reliability of speaker identification for this linguistic group.

**Xing, Li, and Tan (2012)** introduced an innovative approach to speaker identification through the presentation of a hierarchical fuzzy method. The foundation of their method was rooted in the integration of Fuzzy C-Means (FCM) and Fuzzy Support Vector Machine (FSVM) techniques [64]. The overarching goal of the study was to enhance the accuracy of speaker identification by incorporating advanced fuzzy logic techniques into the identification process. The hierarchical fuzzy approach proposed by Xing, Li, and Tan allowed for a more nuanced and flexible representation of speaker characteristics. Fuzzy logic, with its ability to handle uncertainties and imprecise information, provided a suitable framework for capturing the inherent complexity and variability in speaker-specific features. FCM, a clustering algorithm, was employed to delineate distinct speaker clusters, while FSVM, a fuzzy extension of the Support Vector Machine, was utilized for effective classification within these identified clusters. The hierarchical structure of their approach facilitated a multi-level analysis, allowing for a more detailed exploration of speaker characteristics. This hierarchical framework aimed to capture both global and local variations in the feature space, enabling a more comprehensive and accurate representation of individual speaker profiles.

**Srivastava et. al (2013)** delved into the exploration of speaker identification methods based on Granular Fuzzy Models (GFM) [65]. The primary emphasis of their study lay in leveraging granular computing approaches to enhance the efficiency and adaptability of speaker recognition systems. Granular computing involves the manipulation and processing of information at multiple levels of granularity, enabling a more flexible and nuanced analysis of complex data. In the context of speaker identification, Srivastava et al. applied granular computing principles through the use of Granular Fuzzy Models. Fuzzy logic, known for its capacity to handle imprecise and uncertain information, was integrated into the modeling process, allowing for a more robust representation of speaker-specific features. The research aimed to improve the efficiency of speaker recognition systems by incorporating granular computing concepts, which inherently facilitate the management of uncertainty and variability in speech signals. By adopting a granular approach, the researchers sought to capture the intricacies of speaker characteristics at different levels of detail, contributing to a more accurate and adaptive identification process.

**Jawarkar et. al (2013)** extended their exploration of speaker identification, this time with a focused investigation into whispered speech [66]. This study represented a distinctive departure from conventional speech patterns, as it delved into the unique characteristics and nuances associated with the act of whispering. The primary objective was to assess the applicability of whispered speech in the context of reliable speaker recognition. Whispered speech, characterized by a breathy and hushed vocalization, poses specific challenges and opportunities in the domain of speaker identification. Unlike typical spoken communication, whispered speech lacks the vibrancy and regularity found in normal speech, making it a less conventional but potentially valuable modality for speaker recognition systems. The study aimed to unravel the distinctive acoustic features embedded within whispered speech that could serve as robust identifiers for individual speakers. The research methodology likely involved extensive acoustic analysis, examining parameters such as pitch, intensity, and spectral characteristics unique to whispered speech. By comprehensively understanding the acoustic fingerprints of whispered speech, the study sought to determine the feasibility of utilizing this modality for reliable speaker recognition across diverse scenarios.

**Shen et al. (2014)** put forth a novel speaker recognition algorithm that was rooted in the principles of factor analysis, with the overarching goal of enhancing the discriminatory power of speaker features [67]. The primary focus of their study was to contribute insights into the application of factor analysis techniques for the purpose of refining and improving speaker identification processes. Factor analysis, a statistical method commonly used in signal processing, seeks to uncover the underlying factors that contribute to observed variability in data. In the context of speaker recognition, Shen et al. employed factor analysis to dissect the complex interplay of features within speech signals, aiming to identify latent factors that are particularly indicative of individual speaker characteristics. The proposed algorithm likely involved the extraction and analysis of various acoustic features, such as pitch, formants, and spectral characteristics, using factor analysis to uncover latent factors that could serve as discriminative features for speaker identification. By discerning these latent factors, the researchers aimed to enhance the accuracy and reliability of speaker recognition systems.

**Chougule and Chavan (2015)** made a notable contribution to the field of automatic speaker recognition by introducing robust spectral features designed to address mismatch conditions [68]. The study was specifically geared towards enhancing the reliability of speaker identification in real-world scenarios characterized by variations in environmental conditions. The focus of their investigation was on the development of spectral features that could withstand mismatches caused by changes in the acoustic environment. Environmental factors such as background noise, room acoustics, and other external interferences can significantly impact the accuracy of speaker recognition systems. Chougule and Chavan aimed to mitigate these challenges by devising spectral features that were robust and resilient to such variations. The research likely involved the exploration and optimization of various spectral feature extraction techniques, such as Mel-frequency cepstral coefficients (MFCCs) or other representations of the speech signal spectrum. The goal was to identify features that remained consistent and informative across different environmental conditions, ensuring a more reliable and accurate speaker identification process.

**Li et.al (2015)** presented a significant advancement in the realm of speaker recognition by proposing improved deep speaker feature learning specifically tailored for text-dependent speaker recognition [69]. The focus of their research was on leveraging state-of-the-art deep learning techniques to extract more discriminative features from the speech signal, with the ultimate aim of enhancing the accuracy of speaker recognition in text-dependent scenarios. Deep learning, particularly through neural network architectures, has demonstrated remarkable capabilities in learning hierarchical representations of complex data. In the context of speaker recognition, Li et al. sought to harness the power of deep learning to automatically extract high-level features from speech signals that are most relevant for distinguishing between different speakers. The emphasis on text-dependent scenarios highlighted the importance of capturing speaker-specific characteristics associated with particular phrases or words.

**Desai and Tahilramani (2016)** conducted pioneering research that delved into the realm of digital speech watermarking to ensure the authenticity of speakers within a speaker recognition system [70]. Recognizing the critical need for securing speaker identification systems against tampering and unauthorized access, their study proposed a novel approach that integrated digital watermarking techniques into the domain of speaker recognition. Digital speech watermarking involves embedding imperceptible and unique patterns or information directly into the speech signal. In the context of speaker identification, this innovative technique aimed to serve as a robust safeguard against tampering, ensuring the integrity and authenticity of the identified speakers. By incorporating digital speech watermarking, the researchers sought to fortify speaker recognition systems, making them more resilient to malicious attempts to manipulate or forge speaker identities. This research not only addressed a critical aspect of security in speaker recognition technology but also offered a proactive solution to enhance the reliability of speaker identification systems in real-world applications where data integrity is paramount.

**Soleymanpour and Marvi's (2017)** research delved into text-independent speaker identification, with a specific focus on selecting the most similar feature vectors [71]. This study aimed to refine speaker identification methods, particularly in scenarios where spoken content is not predetermined, contributing to the development of more robust text-independent speaker recognition systems. By investigating the selection of comparable feature vectors derived from speech signals, the research sought to enhance the accuracy of matching individual speakers solely based on their voice characteristics. This work addressed a critical aspect of speaker recognition technology, advancing methods for reliable identification in diverse and unpredictable speech contexts

**Zergat et. al (2018)** directed their research towards the optimization of feature selection techniques specifically tailored for G. 729 synthesized speech, aiming to enhance the efficiency and accuracy of automatic speaker recognition systems [72]. G. 729 is a widely used voice compression standard, and optimizing feature selection for synthesized speech is crucial for improving the performance of speaker recognition technologies. The study likely involved exploring various feature selection methods that are well-suited for G. 729 encoded speech signals. By tailoring the feature selection process to the characteristics of synthesized speech, the researchers sought to improve the overall effectiveness of automatic speaker recognition systems operating in environments where G. 729 compression is utilized. This work contributed to the ongoing efforts in optimizing speaker recognition technologies for diverse speech conditions, particularly those involving compressed speech signals.

**Chung et. al (2018)** significantly contributed to the field of speaker recognition by introducing advancements in deep learning techniques, with a specific focus on deep speaker recognition [73]. Their research likely explored novel approaches within the realm of deep learning, which has proven to be a powerful tool for extracting intricate patterns and representations from complex data. In the context of speaker recognition, deep learning methods, such as deep neural networks, have demonstrated the ability to automatically learn discriminative features directly from raw audio signals. By advancing the state-of-the-art in deep speaker recognition, their research potentially played a pivotal role in improving the accuracy and robustness of speaker identification systems, particularly in scenarios where large amounts of data are available for training complex neural networks. This contribution marks a significant step forward in harnessing the capabilities of deep learning for enhancing the capabilities of speaker recognition technology.

**Villalba et. al (2019)** presented a comprehensive overview of the latest developments in speaker recognition technology, with a particular emphasis on highlighting state-of-the-art achievements in the realms of telephone and video speech [74]. The research likely involved an exploration of cutting-edge methodologies and technologies that have advanced the accuracy and efficiency of speaker recognition systems, especially in contexts involving telephony and video communication. This work may have included advancements in feature extraction, pattern recognition algorithms, and the integration of machine learning techniques optimized for handling the challenges unique to telephone and video speech.

**Kelly et. al (2019)** conducted pioneering research by exploring the application of deep neural networks in forensic automatic speaker recognition, specifically utilizing x-vectors [75]. X-vectors represent high-dimensional embeddings generated by deep neural networks, capturing intricate speaker characteristics from speech signals. The research likely delved into the design and optimization of deep neural network architectures tailored for extracting x-vectors and subsequently applying them to forensic scenarios. Forensic automatic speaker recognition is a critical domain where accuracy and reliability are paramount, as it involves the identification of speakers in legal contexts or forensic investigations.

**Jagiasi et. al (2019)** conducted a noteworthy investigation into the application of Convolutional Neural Networks (CNN) for speaker recognition within small-scale systems, specifically addressing concerns related to language and text independence [76]. CNNs, renowned for their effectiveness in extracting hierarchical features from data, were likely employed to automatically learn discriminative patterns from speech signals, eliminating the need for explicit linguistic information. The research entailed the development and optimization of CNN architectures tailored to the constraints of small-scale systems, emphasizing efficiency without compromising recognition accuracy.

**Li et. al (2020)** conducted a study that delved into the development of speaker recognition models tailored for scenarios characterized by limited training data [77]. Recognizing the challenges posed by insufficient data for training robust models, their research likely focused on designing methodologies and algorithms capable of effectively learning and representing speaker characteristics in situations with sparse training samples.

**Jahangir et al. (2020)** delved into the realm of text-independent speaker identification by investigating the synergistic potential of feature fusion and deep neural networks. The integration of these two crucial elements marked a significant advancement in the field, as it aimed to enhance the accuracy and reliability of speaker identification systems. Feature fusion involves the amalgamation of diverse sets of features derived from audio signals, enabling a more comprehensive representation of the speaker's unique characteristics. Meanwhile, deep neural networks, with their ability to automatically extract hierarchical features, provided a sophisticated framework for learning intricate patterns within the data. By combining these techniques, Jahangir et al. sought to address the challenges associated with text-independent speaker identification, where the system must identify and authenticate speakers without relying on specific textual prompts. The outcome of their research not only contributed to the refinement of speaker identification technologies but also opened new avenues for the development of more robust and adaptable systems in the broader domain of audio processing and biometrics.

**Bharath's (2020)** innovative contribution to the field of speaker identification lies in the proposition of an Extreme Learning Machine (ELM) approach tailored to tackle the challenges posed by limited datasets. In the realm of speaker identification, where the availability of extensive and diverse datasets can be a bottleneck for model training and performance, Bharath's work addresses a critical issue. The Extreme Learning Machine, known for its efficacy in handling limited data scenarios, is employed to enhance the robustness and generalization capabilities of speaker identification systems. By leveraging the unique strengths of ELM, such as its ability to rapidly learn and adapt to complex patterns within limited datasets, Bharath's approach opens up possibilities for more practical and real-world applications of speaker identification technology.

**Kabir et al. (2021)** undertook a comprehensive survey dedicated to the expansive domain of speaker recognition. By synthesizing a wealth of knowledge, the researchers delivered an in-depth overview encompassing fundamental theories, diverse recognition methods, and the contemporary opportunities that define the landscape of speaker recognition. This survey serves as a valuable resource for both researchers and practitioners in the field, offering a consolidated understanding of the theoretical underpinnings and methodological advancements in speaker recognition technology.

**Vaessen and Van Leeuwen (2022)** made a significant contribution to the field of speaker recognition by presenting their research, which focused on the fine-tuning of the wav2vec2 model. The wav2vec2 model, renowned for its prowess in acoustic signal processing, serves as the foundation for their investigation. In their study, the researchers sought to enhance the performance of speaker recognition systems by fine-tuning this advanced model, aiming for more accurate and robust results. By delving into the specifics of acoustic signal processing, Vaessen and Van Leeuwen aimed to unlock potential improvements in the model's ability to discern unique speaker characteristics from audio data.

**Table 3: Comparison of Modeling Techniques**

| Study | Approach/Method | Accuracy |
|---|---|---|
| Jawarkar et.al (2011) | Fuzzy Min-Max Neural Network for Speaker Identification | 99.9% |
| Nakagawa et.al (2011) | MFCC and Phase Information for Speaker Identification and Verification | 98.8% |
| Krishnamoorthy et.al (2011) | Introducing Noise for Speaker Recognition under Limited Data | 80% |
| Tolba (2011) | CHMM-Based Approach for Text-Independent Arabic Speaker Identification | 80% |
| Xing et.al (2012) | Hierarchical Fuzzy Speaker Identification with FCM and FSVM | 98.76% |
| Srivastava et.al (2013) | Granular Fuzzy Models for Speaker Identification | 93% |
| Jawarkar et.al (2013) | Whispers Speech for Speaker Identification | 98.6% |
| Shen et.al (2014) | Factor Analysis-Based Speaker Recognition Algorithm | 82.94% |
| Chougule and Chavan (2015) | Robust Spectral Features for Automatic Speaker Recognition | 98% |
| Li et.al (2015) | Improved Deep Speaker Feature Learning for Text-Dependent Recognition | 2% EER |
| Desai and Tahilramani (2016) | Digital Speech Watermarking for Authenticity in Speaker Recognition | 93.33% |
| Soleymanpour and Marvi (2017) | Text-Independent Speaker Identification based on Feature Vectors | 93.2% |
| Zergat et.al (2018) | Feature Selection for G. 729 Synthesized Speech | 0.91% EER |
| Chung et.al (2018) | Advancements in Deep Speaker Recognition | 3.95% EER |
| Villalba et al. (2019) | State-of-the-Art Speaker Recognition for Telephone and Video Speech | 4.95% EER |
| Kelly et al. (2019) | Forensic Automatic Speaker Recognition using x-vectors | 1.40% EER |
| Jagiasi et al. (2019) | CNN-based Speaker Recognition in Small-Scale Systems | 75.2% |
| Li et al. (2020) | Speaker Recognition with Limited Data | 95.0% |
| Jahangir et al. (2020) | Feature Fusion and Deep Neural Network for Text-Independent Recognition | 92.9% |
| Bharath (2020) | ELM Speaker Identification for Limited Dataset | 97.52% |
| Kabir et al. (2021) | Comprehensive Survey on Speaker Recognition | - |
| Vaessen and Van Leeuwen (2022) | Fine-Tuning Wav2vec2 for Speaker Recognition | 1.69% EER |
| Guo et al. (2023) | Dung Beetle Optimized CNN for Speaker Recognition | 97.93% |

**Guo et al. (2023)** have introduced a groundbreaking approach to speaker recognition by unveiling a novel methodology incorporating a Dung Beetle Optimized Convolutional Neural Network (CNN). This innovative research represents a noteworthy stride in the quest for improved recognition accuracy within the realm of speaker identification. The integration of the Dung Beetle Optimization algorithm, inspired by the natural behaviors of dung beetles, with a Convolutional Neural Network underscores the researchers' creative and bio-inspired approach to solving complex problems in pattern recognition. The utilization of the CNN architecture indicates a sophisticated framework for automatically extracting hierarchical features from audio data, while the Dung Beetle Optimization algorithm introduces an element of nature-inspired optimization to enhance the model's performance. The summary of the notable methods is presented in Table 3.

## 8. Conclusion

In conclusion, the in-depth analysis presented in this comprehensive review paper offers a valuable and insightful exploration into the recent strides made in the realm of "Advancements in Real Voice Recognition and Authentication." The paper emphasizes the growing importance of precise identification and authentication of individuals through their actual voices, particularly in the context of diverse applications such as security systems, user identification protocols, and access control mechanisms. A significant strength of this review lies in its meticulous examination of the state-of-the-art developments across various dimensions of real voice recognition. The exploration encompasses cutting-edge advancements in speech recognition algorithms, delves into the intricacies of acoustic-phonetic approaches, scrutinizes the efficacy of pattern recognition methodologies, assesses the utility of template-based techniques, and evaluates the robustness of statistical models and stochastic approaches when applied to real voice recognition scenarios. This multifaceted approach allows for a comprehensive understanding of the diverse strategies employed in the pursuit of accurate and reliable voice-based identification and authentication.

## References

1. Colton, Raymond H., and Jo A. Estill. "Elements of voice quality: perceptual, acoustic, and physiologic aspects." In *Speech and Language*, vol. 5, pp. 311-403. Elsevier, 1981.
2. Johnson, Keith, and Matthias J. Sjerps. "Speaker normalization in speech perception." *The handbook of speech perception* (2021): 145-176.
3. Kathuria, R., Wadehra, A., & Kathuria, V. (2020). Human-centered artificial intelligence: antecedents of trust for the usage of voice biometrics for driving contactless interactions. In *HCI International 2020–Late Breaking Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22* (pp. 325-334). Springer International Publishing.
4. Jain, Anil K., Arun Ross, and Salil Prabhakar. "An introduction to biometric recognition." *IEEE Transactions on circuits and systems for video technology* 14, no. 1 (2004): 4-20.
5. Kambeyanda, Dona, Lois Singer, and Stan Cronk. "Potential problems associated with use of speech recognition products." *Assistive Technology* 9, no. 2 (1997): 95-101.
6. Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano. "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds." *Applied Sciences* 6, no. 5 (2016): 143.
7. Togneri, Roberto, and Daniel Pullella. "An overview of speaker identification: Accuracy and robustness issues." *IEEE circuits and systems magazine* 11, no. 2 (2011): 23-61.
8. Bimbot, Frédéric, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. "A tutorial on text-independent speaker verification." *EURASIP Journal on Advances in Signal Processing* 2004 (2004): 1-22.
9. MacNeilage, Peter F. *The origin of speech*. No. 10. Oxford University Press, 2010.
10. Yu, Dong, and Lin Deng. *Automatic speech recognition*. Vol. 1. Berlin: Springer, 2016.
11. Khadse, Kavita. "To Study And Analyze The Use Of Speech Recognition Systems And Its Benefits Across Various Demographics."
12. Mian Qaisar, Saeed. "Isolated speech recognition and its transformation in visual signs." *Journal of Electrical Engineering & Technology* 14 (2019): 955-964.
13. Schultz, Stefan. "Hello, computer. Approaches to designing speech-based user experiences." (2018).

14. Gaspar, Laszlo. *Assessment of signal preprocessing for speech recognition*. Nottingham Trent University (United Kingdom), 1996.
15. Bömers, Florian. "Wavelets in real time digital audio processing: Analysis and sample implementations." *University of Manheim, Dept. of Computer Science IV, Master's thesis* (2000).
16. Spanias, Andreas, Ted Painter, and Venkatraman Atti. *Audio signal processing and coding*. John Wiley & Sons, 2006.
17. Johnson, Keith, and Matthias J. Sjerps. "Speaker normalization in speech perception." *The handbook of speech perception* (2021): 145-176.
18. Hagmüller, Martin. *Speech enhancement for disordered and substitution voices*. na, 2009.
19. Shannon, Robert V., Fan-Gang Zeng, and John Wygonski. "Speech recognition with altered spectral distribution of envelope cues." *The Journal of the Acoustical Society of America* 104, no. 4 (1998): 2467-2476.
20. Ismail, Muhammad, Shahzad Memon, Lachhman Das Dhomeja, Shahid Munir Shah, Dostdar Hussain, Sabit Rahim, and Imran Ali. "Development of a regional voice dataset and speaker classification based on machine learning." *Journal of Big Data* 8 (2021): 1-18.
21. Vipperla, Ravichander. "Automatic Speech Recognition for ageing voices." (2011).
22. Hamid, Oday Kamil. "Frame blocking and windowing speech signal." *Journal of Information, Communication, and Intelligence Systems (JICIS)* 4, no. 5 (2018): 87-94.
23. Trivedi, Nitin, Vikesh Kumar, Saurabh Singh, Sachin Ahuja, and Raman Chadha. "Speech recognition by wavelet analysis." *International Journal of Computer Applications* 15, no. 8 (2011): 27-32.
24. Zhang, Shucong, Erfan Loweimi, Peter Bell, and Steve Renals. "Windowed attention mechanisms for speech recognition." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7100-7104. IEEE, 2019.
25. Sahidullah, Md, and Goutam Saha. "A novel windowing technique for efficient computation of MFCC for speaker recognition." *IEEE signal processing letters* 20, no. 2 (2012): 149-152.
26. Liu, Weiqiang, Qicong Liao, Fei Qiao, Weijie Xia, Chenghua Wang, and Fabrizio Lombardi. "Approximate designs for fast Fourier transform (FFT) with application to speech recognition." *IEEE Transactions on Circuits and Systems I: Regular Papers* 66, no. 12 (2019): 4727-4739.
27. Hopper, Greg, and Reza Adhami. "An FFT-based speech recognition system." *Journal of the Franklin Institute* 329, no. 3 (1992): 555-562.
28. Raj, Bhiksha, Lorenzo Turicchia, Bent Schmidt-Nielsen, and Rahul Sarpeshkar. "An FFT-based companding front end for noise-robust automatic speech recognition." *EURASIP Journal on Audio, Speech, and Music Processing* 2007 (2007): 1-13.
29. Kinnunen, Tomi, and Haizhou Li. "An overview of text-independent speaker recognition: From features to supervectors." *Speech communication* 52, no. 1 (2010): 12-40.
30. Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review." *IEEE Signal processing magazine* 32, no. 6 (2015): 74-99.
31. Wu, Jian-Da, and Bing-Fu Lin. "Speaker identification based on the frame linear predictive coding spectrum technique." *Expert Systems with Applications* 36, no. 4 (2009): 8056-8063.
32. Farah, Shahzadi, and Azra Shamim. "Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization." In *2013 3rd IEEE International Conference on Computer, Control and Communication (IC4)*, pp. 1-5. IEEE, 2013.
33. Todkar, Satyam P., Snehal S. Babar, Rudrendra U. Ambike, Prasad B. Suryakar, and J. R. Prasad. "Speaker recognition techniques: A review." In *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1-5. IEEE, 2018.
34. Li, Penghua, Fangchao Hu, Yinguo Li, and Yang Xu. "Speaker identification using linear predictive cepstral coefficients and general regression neural network." In *Proceedings of the 33rd Chinese Control Conference*, pp. 4952-4956. IEEE, 2014.
35. Farah, Shahzadi, and Azra Shamim. "Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization." In *2013 3rd IEEE International Conference on Computer, Control and Communication (IC4)*, pp. 1-5. IEEE, 2013.
36. Sandhya, P., V. Spoorthy, Shashidhar G. Koolagudi, and N. V. Sobhana. "Spectral features for emotional speaker recognition." In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pp. 1-6. IEEE, 2020.

37. Yujin, Yuan, Zhao Peihua, and Zhou Qun. "Research of speaker recognition based on combination of LPCC and MFCC." In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 3, pp. 765-767. IEEE, 2010.
38. Chauhan, Paresh M., and Nikita P. Desai. "Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using wiener filter." In *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, pp. 1-5. IEEE, 2014.
39. Wong, Eddie, and Sridha Sridharan. "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification." In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*, pp. 95-98. IEEE, 2001.
40. Jing, Xinxing, Jinlong Ma, Jing Zhao, and Haiyan Yang. "Speaker recognition based on principal component analysis of LPCC and MFCC." In *2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 403-408. IEEE, 2014.
41. Reynolds, Douglas Alan. *A Gaussian mixture modeling approach to text-independent speaker identification*. Georgia Institute of Technology, 1992.
42. Kumar, G. Suvarna, KA Prasad Raju, Mohan Rao CPVNJ, and P. Satheesh. "Speaker recognition using GMM." *International Journal of Engineering Science and Technology* 2, no. 6 (2010): 2428-2436.
43. Burget, Lukas, Pavel Matejka, Petr Schwarz, Ondrej Glembek, and Jan Honza Cernocky. "Analysis of feature extraction and channel compensation in a GMM speaker recognition system." *IEEE Transactions on Audio, Speech, and Language Processing* 15, no. 7 (2007): 1979-1986.
44. Ding, Ing-Jr, Chih-Ta Yen, and Da-Cheng Ou. "A method to integrate GMM, SVM and DTW for speaker recognition." *International Journal of Engineering and Technology Innovation* 4, no. 1 (2014): 38-47.
45. Gales, Mark, and Steve Young. "The application of hidden Markov models in speech recognition." *Foundations and Trends® in Signal Processing* 1, no. 3 (2008): 195-304.
46. Inman, Michael, Douglas Danforth, S. Hangai, and K. Sato. "Speaker identification using hidden Markov models." In *ICSP'98. 1998 Fourth International Conference on Signal Processing (Cat. No. 98TH8344)*, pp. 609-612. IEEE, 1998.
47. Sha, Fei, and Lawrence Saul. "Large margin hidden Markov models for automatic speech recognition." *Advances in neural information processing systems* 19 (2006).
48. Farrell, Kevin R., Richard J. Mammone, and Khaled T. Assaleh. "Speaker recognition using neural networks and conventional classifiers." *IEEE Transactions on speech and audio processing* 2, no. 1 (1994): 194-205.
49. Liu, Zheli, Zhendong Wu, Tong Li, Jin Li, and Chao Shen. "GMM and CNN hybrid method for short utterance speaker recognition." *IEEE Transactions on Industrial informatics* 14, no. 7 (2018): 3244-3252.
50. Saon, George, Zoltán Tüske, Daniel Bolanos, and Brian Kingsbury. "Advancing RNN transducer technology for speech recognition." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5654-5658. IEEE, 2021.
51. Sztahó, Dávid, György Szaszák, and András Beke. "Deep learning methods in speaker recognition: a review." *arXiv preprint arXiv:1911.06615* (2019).
52. Soong, Frank K., Aaron E. Rosenberg, Bling-Hwang Juang, and Lawrence R. Rabiner. "Report: A vector quantization approach to speaker recognition." *AT&T technical journal* 66, no. 2 (1987): 14-26.
53. Bharti, Roma, and Priyanka Bansal. "Real time speaker recognition system using MFCC and vector quantization technique." *International Journal of Computer Applications* 117, no. 1 (2015).
54. Gupta, Arnav, and Harshit Gupta. "Applications of MFCC and Vector Quantization in speaker recognition." In *2013 International conference on intelligent systems and signal processing (ISSP)*, pp. 170-173. IEEE, 2013.
55. Martinez, Jorge, Hector Perez, Enrique Escamilla, and Masahisa Mabo Suzuki. "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques." In *Conielecomp 2012, 22nd International conference on electrical communications and computers*, pp. 248-251. IEEE, 2012.
56. Campbell, William M., Joseph P. Campbell, Terry P. Gleason, Douglas A. Reynolds, and Wade Shen. "Speaker verification using support vector machines and high-level features." *IEEE Transactions on Audio, Speech, and Language Processing* 15, no. 7 (2007): 2085-2094.

57. Ismail, Ahmed, Samir Abdlerazek, and Ibrahim M. El-Henawy. "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping." *Sustainability* 12, no. 6 (2020): 2403.
58. Chauhan, Neha, Tsuyoshi Isshiki, and Dongju Li. "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database." In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 130-133. IEEE, 2019.
59. Kanisha, B., S. Lokesh, Priyan Malarvizhi Kumar, P. Parthasarathy, and Gokulnath Chandra Babu. "Speech recognition with improved support vector machine using dual classifiers and cross fitness validation." *Personal and ubiquitous computing* 22 (2018): 1083-1091.
60. Jawarkar, N. P., R. S. Holambe, and T. K. Basu. "Use of fuzzy min-max neural network for speaker identification." In *2011 International conference on recent trends in information technology (ICRTIT)*, pp. 178-182. IEEE, 2011.
61. Nakagawa, Seiichi, Longbiao Wang, and Shinji Ohtsuka. "Speaker identification and verification by combining MFCC and phase information." *IEEE transactions on audio, speech, and language processing* 20, no. 4 (2011): 1085-1095.
62. Krishnamoorthy, P., H. S. Jayanna, and SR Mahadeva Prasanna. "Speaker recognition under limited data condition by noise addition." *Expert Systems with Applications* 38, no. 10 (2011): 13487-13490.
63. Tolba, Hesham. "A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach." *Alexandria Engineering Journal* 50, no. 1 (2011): 43-47.
64. Xing, YuJuan, Hengjie Li, and Ping Tan. "Hierarchical fuzzy speaker identification based on FCM and FSVM." In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 311-315. IEEE, 2012.
65. Bhardwaj, Saurabh, Smriti Srivastava, Madasu Hanmandlu, and J. R. P. Gupta. "GFM-based methods for speaker identification." *IEEE transactions on cybernetics* 43, no. 3 (2013): 1047-1058.
66. Jawarkar, Naresh P., Raghunath S. Holambe, and Tapan Kumar Basu. "Speaker identification using whispered speech." In *2013 International Conference on Communication Systems and Network Technologies*, pp. 778-781. IEEE, 2013
67. Shen, Xuanjing, Yujie Zhai, Yu Wang, and Haipeng Chen. "A speaker recognition algorithm based on factor analysis." In *2014 7th International Congress on Image and Signal Processing*, pp. 897-901. IEEE, 2014.
68. Chougule, Sharada V., and Mahesh S. Chavan. "Robust spectral features for automatic speaker recognition in mismatch condition." *Procedia Computer Science* 58 (2015): 272-279.
69. Li, Lantian, Yiye Lin, Zhiyong Zhang, and Dong Wang. "Improved deep speaker feature learning for text-dependent speaker recognition." In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 426-429. IEEE, 2015.
70. Desai, Nihalkumar, and Nikunj Tahilramani. "Digital speech watermarking for authenticity of speaker in speaker recognition system." In *2016 international conference on micro-electronics and telecommunication engineering (ICMETE)*, pp. 105-109. IEEE, 2016.
71. Soleymanpour, Mohammad, and Hossein Marvi. "Text-independent speaker identification based on selection of the most similar feature vectors." *International Journal of Speech Technology* 20 (2017): 99-108.
72. Zergat, Kawthar Yasmine, Sid-Ahmed Selouani, and Abderrahmane Amrouche. "Feature Selection Applied to G. 729 Synthesized Speech for Automatic Speaker Recognition." In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pp. 178-182. IEEE, 2018.
73. Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." *arXiv preprint arXiv:1806.05622* (2018).
74. Villalba, Jesús, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom et al. "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18." In *Interspeech*, pp. 1488-1492. 2019.
75. Kelly, Finnian, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander. "Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors." In *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*. Audio Engineering Society, 2019.

76. Jagiasi, Rohan, Shubham Ghosalkar, Punit Kulal, and Asha Bharambe. "CNN based speaker recognition in language and text-independent small scale system." In *2019 third international conference on i-smac (iot in social, mobile, analytics and cloud)(I-SMAC)*, pp. 176-179. IEEE, 2019.
77. Li, Ruirui, Jyun-Yu Jiang, Jiahao Liu, Chu-Cheng Hsieh, and Wei Wang. "Automatic speaker recognition with limited data." In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 340-348. 2020.
78. Jahangir, Rashid, Ying Wah Teh, Nisar Ahmed Memon, Ghulam Mujtaba, Mahdi Zareei, Uzair Ishtiaq, Muhammad Zaheer Akhtar, and Ihsan Ali. "Text-independent speaker identification through feature fusion and deep neural network." *IEEE Access* 8 (2020): 32187-32202.
79. KP, Bharath. "ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score." *Multimedia Tools and Applications* 79 (2020): 28859-28883.
80. Kabir, Muhammad Mohsin, Muhammad F. Mridha, Jungpil Shin, Israt Jahan, and Abu Quwsar Ohi. "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities." *IEEE Access* 9 (2021): 79236-79263.
81. Vaessen, Nik, and David A. Van Leeuwen. "Fine-tuning wav2vec2 for speaker recognition." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7967-7971. IEEE, 2022.
82. Guo, Xinhua, Xiao Qin, Qing Zhang, Yuanhuai Zhang, Pan Wang, and Zhun Fan. "Speaker Recognition Based on Dung Beetle Optimized CNN." *Applied Sciences* 13, no. 17 (2023): 9787.