**Khushboo Jha[1]*,**

**Aruna Jain**
**Sumit Srivastava**

# Analysis of Human Voice for Speaker Recognition: Concepts and Advancement


**Abstract:**
Human voice or speech is a contactless, non-invasive biometric trait for human recognition, easy to use with minimal computer complexity and inexpensive to implement. Speaker recognition (SR) has turned out to be a magnificent approach using speech as the central premise since decades. Its broad range of usages, like forensic speech verification to identify culprits by law enforcement authorities and access control to mobile banking, mobile shopping, etc., has made it a lucrative area of research. Also, the ease of use and dependability of SR will significantly assist people with disabilities in securely accessing and reaping the benefits of digital-era services. Additionally, the emergence of numerous deep learning methods for feature extraction and classification, has helped SR to achieve tremendous progress. This paper presents a comprehensive study on the progression of SR for decades till the present, including integration with Blockchain and challenges. It covers most of the factors that influence SR performance such as fundamentals and structure of SR, different speech pre-processing techniques, various speech features, feature extraction techniques, traditional and neural network-based classification techniques and deep learning-based SR toolkits. As a consequence, in this digital Blockchain era, it will help to design robust and reliable recognition-based services for mankind.

**Keywords:** acoustic signal processing, feature extraction, human voice, speech analysis, neural networks, blockchain

## 1. INTRODUCTION

As a consequence of the growing intrigue in security and the significant reliance on internet usage in the epoch of the COVID-19 (corona virus disease 2019) pandemic, biometric utilisation has increased. The biometric alternative can be used to circumvent the breaches of the password-based system. Moreover, speech being a physiological as well as behavioural biometric trait [1] is the most natural, non-invasive, contactless, widely accepted and hence fundamental way of distinguishing a person among all the biometric traits. Speech features include pace, volume, pitch level, and quality, whereas speaking style determines articulation and pauses. A person's voice transmits the information used to identify someone's identity. Each person has an individual voice and this makes each person distinguishable. The accent indicates the speaker's attributes, such as age, race, gender, etc. SR is the technique of recognizing individuals based on their acoustic features of speech. Because of physiological (e.g., vocal tract shapes, larynx sizes, etc.) and behavioural (e.g., speaking style, pronunciation pattern, vocabulary choice, etc.) variations in the speech production system, no two people sound alike. Also, people speak in different ways, resulting in disparities in speaking rate, accent, intonation, etc. SR systems strive to use these variables to distinguish among speakers.

Some of the human recognition is difficult for disabled people for access to various services [2] such as banking, security, etc. as they are sometimes unable to access the instruments of transaction or devices like physical signature, ATM, touch screen interface, biometric devices, etc. On the technology front, compared to other modalities such as hand geometry, fingerprint, face recognition, etc., which requires greater user cooperation and certain proximity to the device, SR is relatively convenient and reliable considering the ease in accessing speech receiving device and the contactless nature of this recognition technique. This is equally relevant in the case of accessibility for disabled children.

Moreover, SR is used in a variety of applications [3], such as authentication, forensic speech verification, biometric, security and surveillance, mobile banking, in-car systems, education, air traffic control and even human-computer interaction. Most smart phones have virtual assistant software that uses speech commands to aid systems. Technology advancement permits businesses and consumers to accomplish SR at a low cost. This technology can also be used to substitute touch tone dialing, effective ways to reach customers who speak a variety of languages. Table I represents the evolution of SR [3-9] since decades till date. Also, due to the following reasons, SR is the most popular biometric technology for human-machine interaction:

- The only technology that processes acoustic information, in contrast with other recognition system, which usually use image information as in case of face, iris, fingerprint, etc.

- When someone speaks, it's unnecessary to do anything exceptional to recognise them as the speaker. There are typical privacy concerns when using SR for surveillance or in general, when the target isn't aware of it.
- Non-obtrusive technology, as the equipment required to collect speech samples just need a simple microphone.
- Areas such as verifying speech-based transactions i.e., telebanking, ATM and forensics where the results could be used as evidence in court cases, etc., are the perk of speaker recognition.
- Speech based systems can equip a smart home with a SR system that includes features to assist visually challenged and senior citizens in controlling devices.

Moreover, an advanced SR system works better than ever for verification and identification of the speech command. Therefore, it is considered one of the handiest biometric features of an individual's identity.

**Table I**. Progression of Speaker Recognition from 1960s till Present

| Year | Progress |
|---|---|
| 1960-1969 | Lawrence G. Kersta, the Bell Laboratories Physicist introduced first SR system.<br>The frequency domain was used to analyse the spectrum for various features such as pitch, formants, and so on.<br>Fast Fourier transform was developed. |
| 1970-1979 | Text-dependent speaker verification system was developed.<br>Support Vector Machine (SVM) was introduced.<br>Audio feature extraction technique: Joint time-frequency technique. |
| 1980-1989 | Score normalization and cepstrum based systems was introduced.<br>HMM based text-dependent methods was developed.<br>Vector Quantization codebooks. |
| 1990-1999 | First National Institute of Standards and Technology (NIST), Speaker recognition evaluation (SRE96).<br>SR method shifted from template based pattern matching to statistical modeling (e.g. GMM, GMM-UBM, etc.) |
| 2000-2009 | The Super SID project exploited high-level features for high-accuracy SR.<br>HMM was employed with the feed-forward Artificial Neural Network (ANN).<br>Most research focussed on compensation of mismatches (methods like i-vector and PLDA) and development of practical speaker verification systems.<br>With the burgeoning of deep learning (DL), deep features was extensively used.<br>First SR toolkit (ALIZE) developed by French research Ministry Technolangue program.<br>SRE used non-English data.<br>NIST added an unsupervised adaptation feature in which systems would optionally upgrade the speaker model.<br>SR through coded speech under matched and mismatched environment was introduced. |
| 2010-present | A human-assisted SR was introduced.<br>Researchers largely focused on different test conditions such as added noise and room reverberation.<br>Fusion of audio-visual data.<br>Exploration of audio extracted from amateur video data.<br>DL approach is preferred.<br>Deep neural network (DNN) based bottleneck (BN) features became popular in text-dependent speaker verification. |

The rest of this paper is structured as follows: Section 2 discusses the fundamentals of SR. Section 3 explores types of features and feature extraction techniques. Section 4 discusses different categories and types of classification methods. Section 5 discusses SR programming toolkits. Section 6 discusses the challenges of SR system. Section 7 concludes the paper.

## 2. Fundamentals of speaker recognition

In 1962, a research "Voiceprint Identification" by Lawrence G. Kersta introduced the first SR system [6], defined as a branch of digital signal processing concerned with recognizing a speaker by their speech features. In the recent few decades, there has been a tremendous growth in SR [8] research due to developments in signal processing, algorithms, architecture and hardware [10]. The basic framework of SR can be divided into three modules such as pre-processing, feature extraction and classifier, discussed below:

*A. Pre-processing*

A microphone is usually used to collect the speech signal sent to SR. This means that noise may also be sent along with the speech. The purpose of speech pre-processing [11] is to lower the signal-to-noise ratio (SNR) and it boost the effectiveness of the subsequent modules such as features extraction, modelling, and feature matching. Consequently, the classification rate and computation time of the system is enhanced. Speech signal is filtered and manipulated to diminish unwanted noise, approaches such as voice activity detection (VAD), pre-emphasis, framing, windowing, energy normalisation, end-point detection, etc. are commonly utilised as discussed below:

- Voice activity detection - VAD [11] splits the speech signal into short frames to determine whether or not each frame contains speech, i.e., identifies the voiced sections of speech. Such methods are built on the selection of features that characterize the distinctive features of noise and speech.
- Pre-emphasis - The high-frequency feature of speech signal is pre-emphasized [11] to flatten the intensity spectrum with the help of a high pass filter and equalize the high and low-frequency aspects. Also, pre-emphasis filter minimize the high dynamic range of speech waveforms and improve the signal-to-noise ratio. Human speech signal has a spectral slope of around -6dB, pre-emphasis is carried out to eliminate this slope. For this purpose, a high-pass finite impulse response (FIR) filter of order 1 is applied to the speech signal as defined below:

$$y[n] = s[n] - P\ s[n-1] \quad (1)$$

Here, y[n] denotes pre-emphasized sample, s[n] denotes nth speech signal and P denotes the pre-emphasis factor, ranging from 0.9 to 1. Pre-emphasis guarantees that every speech signal pattern is of akin amplitude in the frequency domain, which give them identical weightage in the successive processing phases.

- Framing - It is a technique of dividing an ongoing stream of speech signal into blocks of consistent length in order to simplify block-wise processing of the signal. This merely denotes that the transmission is separated or blocked into frames of 20 to 30 milliseconds. As a result, adjacent frames are typically 30 to 50 percent overlapped, in order to ensure that no crucial information from the speech signal is lost due to windowing.
- Windowing - It is an approach [11] of amplifying a waveform of a speech signal fragment by a time window of a specified shape in order to emphasise predetermined signal features. The speech signal must be attenuated to zero or near to zero to reduce the discontinuity at the start and end of every frame, hence lowering the mismatch. Also, windowing a speech signal aims to avoid truncation complications by smoothing the speech signal. A long window with frame lengths of around 20 to 30 milliseconds, having frame space of 5 to 10 milliseconds is preferred for obtaining superior frequency resolution. The selection of a window requires a balance between several competing aspects:

a) Window shape can minimise discrepancies but affect the signal form, as length is directly proportional to frequency resolution but inversely proportional to time resolution.

b) And signal overlap, which is proportional to the frame rate and correlation of successive frames. FIR filter design for speech employs a variety of window functions, including Hanning, Hamming, Rectangular, Triangular, Bartlett, etc. out of which Hamming window is the state-of-the-art. The Hamming window function [12] is represented by:

$$w(n) = 0.54 - 0.46 \cos(2\pi n/(n-1)) , 0 \leq n \leq N \quad (2)$$

Here w(n) is the window function and N is the frame size.

- End-point Detection - The energy levels of speech signal have a high edge after a starting point and a low edge before a finishing point, known as beginning and ending point respectively. To identify endpoints [13], the cepstral feature must adopt low-complexity and short-term energy. These endpoints are automatically aligned to the SR feature vector and processing is minimised from the speech sampling rate to the frame rate. To detect significant endpoint among the accessible endpoints of energy feature, a good detector is required. As the detector output may include erroneous acceptances, a decision unit is needed to make conclusive judgments using detection findings. As a result, at first, we need to identify the edges and then locate their respective endpoints, as endpoints detection is inextricably linked to edges.
- Energy Normalization - Energy normalization is intended to normalize the energy of speech which is accomplished by determining the highest energy level across the spoken phrases.

*B. Feature extraction*

After the speech signal has been cleaned up, it is delivered to the next module for feature extraction. Often termed as front end pre-processing [6] , it is used for training and testing phases of the SR system. Typically, features are a preset number of coefficients or values derived by implementing different techniques to the speech signal which must resist factors like noise and the echo effect. Also, the distinctive feature extracted from the speech signal reveals the speaker's identity. A detailed discussion on classification of features and feature extraction techniques are done in section 3.

*C. Classification*

The extracted features have a critical role in the classification module's performance and efficiency [10]. This module uses the extracted features to generate the SR algorithms for feature matching of the input speech signal, different classification models [15], traditional and neural network based classification techniques are discussed in section 4.

*D. Working of the speaker recognition system*

The SR system must pass through two stages, namely training and testing [6] as shown in Fig. 1.

- Training stage - During this stage, the initial step involves the acquisition of the speech signal from an individual who is enrolling. The initial signal undergoes a pre-processing stage to eliminate any unwanted noise. Subsequently, the refined signal is directed to the feature extraction module, which aims to identify and extract key features that enable differentiation between different entities. This feature vector or template is used to establish the speaker model [15] and then stored in the database.
- Testing stage - The SR system captures the speech signal, extracts feature and compares it against the trained reference template saved in the database. The higher similarity score or threshold between the query and the reference template, determine whether the user is recognised or rejected.

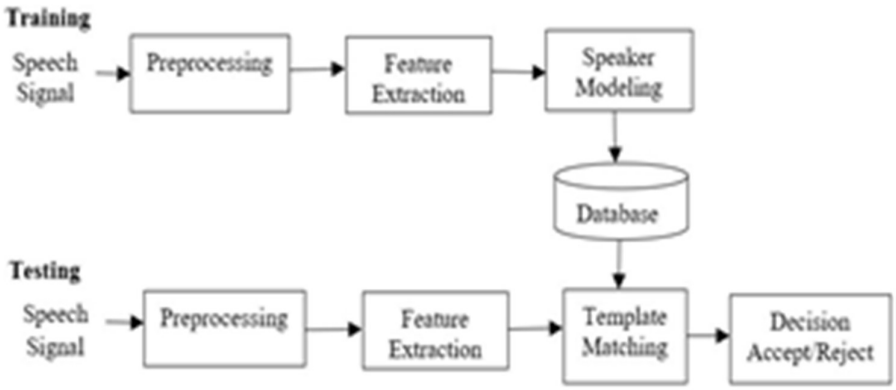



Figure 1. Block diagram of speaker recognition system

The techniques of SR can be classified into two main categories: text-dependent methods and text-independent methods [14-16]. In terms of training and assessment, the text-dependent method necessitates the speaker to articulate a specific word or set of words that are crucial. The text-dependent method uses template i.e., model sequence comparison approach [17]. In this method, the time axes of the speech specimen and reference sample of enrolled speaker are aligned. The similarity between them is observed during the entire duration of the spoken discourse. Such approaches yield better identification results than the text-independent method because it may directly use speech distinctiveness linked with each speaker but at the same time it is vulnerable to spoofing attacks. Text-independent approaches refer to techniques that are not reliant on specific texts throughout both the training and testing phases. Consequently, there is no requirement to recall a password. This feature enhances the system's resilience to spoofing assaults.

*E. Speaker identification and speaker verification*

Based on various tasks, SR is classified as speaker identification (SI) and speaker verification (SV). In order to identify an unknown speaker, SI [14-23] analyses their verbal output. From the set of all N registered speakers, it picks the right one, as shown in Fig. 2. Specifically, we perform N comparisons, each of which involves matching a unique utterance with one of N templates already present in the system. SV [9, 14, 24] examines the speech to determine if the claimed speaker is authentic or a counterfeit. The verification system compares the claimed speaker's speech to the stored trained template of previously registered speaker i.e., it is a 1:1 match as shown in Fig. 2. If both the templates match and exceed a predetermined threshold, the claimed speaker is successfully validated or else rejected. As a result, this is a procedure for determining the credibility of the claimed speaker. Or we can say that the practice of recognising or rejecting a speaker's asserted identification is known as speaker verification. It encompasses most applications in which speech is used to validate a speaker's identification.

In an audio model database, the procured features of the SI system are found to correlate with all of the characteristics of the speakers. In comparison, the obtained features in SV systems are only associated with the registered speaker's stored/saved features [6] that the one asserted to be. SV is used for a wide variety of purposes, such as access control, vetting credit card purchases, authorising money transfers, and verifying phone banking transactions.

## 3. Feature Extraction and classification of speech features

Feature extraction, also known as a front-end technique, is a process that transforms speech samples into a series of feature vector coefficients. These coefficients specifically contain the necessary information to identify a particular utterance. The speech signal encompasses a variety of features [6, 11]. However, just a subset of these features is suitable for distinguishing between speakers. The following are some qualities that should be present in paragon feature:

- High inter speaker disparity and low intra speaker disparity.
- Resistant to distortion and noise.

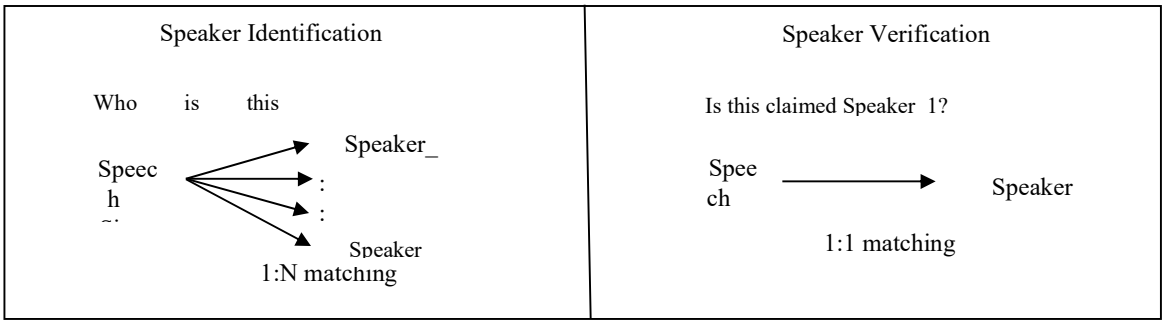



**Figure 2.** Speaker identification and speaker verification

- Transpire naturally in speech.
- Tough to spoofing.
- Must disregard the speaker's age and feelings.

Speech features are higher-level representations than raw data representations, such as frequencies instead of raw temporal samples. SR is a way of recognising people based on their inherent features [6, 8], Fig. 3 represents the classification of speech features and described in details below:

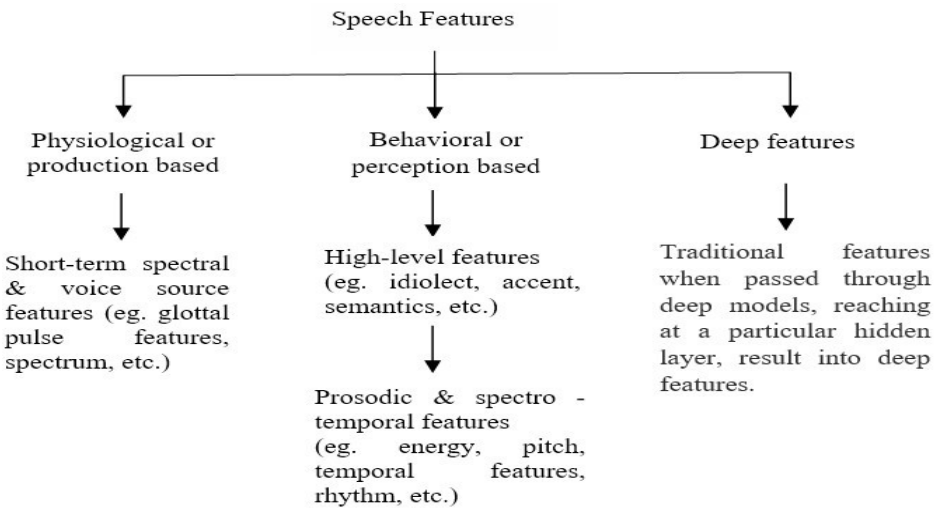



**Figure 3.** Classification of speech features

### *A. Physiological based features*

Such features depend on the vocal tract's dimension, length, fold size, etc. They are easy to extract, requires less data but gets affected by channel mismatch and noise. They are further classified as discussed below:

1. *Short-term spectral features* – It depicts the physical properties of the vocal tract. It is discriminative, easier to compute and improves performance of SR system. The speech signal being time varying signal, need to be analyzed in short frame. The frame duration may lie between 20–30 milliseconds. For the aforesaid time interval, the speech signal is believed to be stationary. Pre-emphasis and windowing are done prior to feature extraction stage. From each frame, a feature vector is extracted. They are further categorised as Spectrum and Gammatone pulse features [6]. Some of the renowned feature extraction techniques are discussed below:

- *Mel-frequency Cepstral Coefficients (MFCC)* – MFCC [18] was introduced in 1980 by Davis and till date preferred by the researchers. It is based on the known variation of the human ear's critical bandwidth with frequency. It represents spectral characteristics of an utterance of speech, which is one of the prominent features comparable to the human auditory perception system. Thus, widely used as the feature extraction technique for SR tasks. MFCC has been the most widely used feature due to high recognition accuracy, lower complexity and the capability to capture a major characteristics of speech signal but the performance is largely affected by the background noise. Feature extraction of MFCC starts with pre-emphasize of input speech signal. Then short-time fourier approach is performed to get the magnitude spectrum. The magnitude spectrum is wrapped into mel-spectrum through triangular overlapping windows in which centre frequency of window is uniformly distributed using mel scale. The log operation on the power spectrum is performed. Lastly, discrete cosine transform (DCT) is applied on the log mel power-spectrum to extract the cepstral features, as shown below.

  $c(n)=\sum_{(m=0)}\hat{}(M-1)[\log(s(m)\cos(n\pi(m-0.5))/M)]$ (3)

  for n=0,1,2,…, C-1. Here, c(n) denotes cepstral coefficients, s(m) is the magnitude spectrum and C is the total number of MFCCs. The coefficients obtained after the applying DCT are considered as MFCC features.
- *Gammatone Frequency Cepstral Coefficients (GFCC)*- One of the key drawbacks of MFCC is its vulnerability to additive noise. The GFCC [20] are a group of Gammatone Filter banks-based auditory features. The outcome of the Gammatone filter bank is known as Cochleagram [25], a frequency-time portrayal of the speech signal. GFCC feature extraction process starts with passing the input speech signal through a gammatone filterbank with 64-channel. As a method of time windowing, thoroughly correct the filter response at each channel (by taking an exact value) and overcome it to 100 Hz. The exact value is then calculated, resulting in a time-frequency (T-F) representation, a cochleagram variant. At last, cube

root of the aforesaid representation is calculated followed by DCT to the derived cepstral features, as shown below.

$$g(n;u)=(2/M)^{0.5}\{\sum_{(i=0)}^{(M-1)}[1/3 \log(\bar{y}(n;i))(\cos(\pi u(2i-1))/2M)]\} ;\ u = 0,1,\ldots,31 \quad (4)$$

The first 12 coefficients result into GFCC features. GFCC is enhanced with the 1st and 2nd order derivatives known as deltas (12 coefficients) and double deltas (12 coefficients), bringing the total GFCC features to 36. The features in GFCC are more resilient than in MFCC because of the cube root rectification.

- *Linear Prediction Cepstral Coefficients (LPCC)* -
  LPCCs are obtained from the Linear Prediction (LP) method [6] of the speech signal. The LP model assumes an all-pole filter to model the vocal tract response. The spectrum obtained from the model is defined below and called as LP spectrum.

$$H(z)=1/(1-\sum_{(i=1)}^{p} a_i z_i^{(-1)}) \quad (5)$$

where ai is the LP filter coefficient and p being the order of the LP model. DCT is performed on the log LP spectrum to get the LP Cepstral Coefficients.1,11 The first few coefficients are retained to form the LPCC feature vector. One of the drawbacks with LPCC is the sensitivity to the LP order. A high LP order gives a spurious peaks in the LP spectrum whereas a low order gives a poor approximation.

2. *Voice source features* – Features like fundamental frequency and glottal pulse shape, characterize the glottal excitation signal of voiced sounds [8] and it is presumed that speaker-specific feature is conveyed. Because of the filtering effect of the vocal tract, it is difficult to directly measure the glottal features.

*B. Behavioral based speech features*

They are derived by evaluating parameters such as socioeconomic status, personality type, education, place of birth, language, etc. They are difficult to extract as they involve large training data but robust against noise and channel effects.

*1. High level features* – High-level features [8] assess many characteristics of an individual's speech signal, such as the utilisation of particular words, accent, and duration of pauses between words, across time intervals that exceed a few seconds. Early explorations in the field of speech recognition (SR) involved the utilisation of high-level features. However, these initial attempts were met with limited success. These features convey conversational-level properties of the speaker, like: the use of a specific word hence can be easily mimicked. Also, speaker's distinct vocabulary (known as idiolect) is utilised to characterize the speaker i.e., pertains to the statistical analysis of word sequences in relation to individual speakers. In nutshell, these features are long time feature, spanning over the time interval of a word or utterance reflecting origin, race, etc. Example: pronunciation, prosody, word duration, etc.

*2. Prosodic features* – It span over lengthy segments such as syllables, utterances, words, etc. It extracts information about the way of speaking and thus it can be easily mimicked. Fundamental frequency (F0) is considered as the expedient prosodic feature. The melding of F0 with spectral features outperforms even in noisy conditions. The prosodic features [8] can be further divided into source features and supra segmental features depending on the time interval of the examined speech part. Source features span over a single glottal period wherein supra segmental features span over some glottal interval. Example: formant, pitch, energy contours, etc.

*3. Spectro-temporal feature* – Speech features perceptible over 100-500 ms gaps are typically gleaned from the signal's spectrogram, thereby the name spectro-temporal features [8]. The spectrogram portrayal is formed by concatenation of frame spectra, resulting in the 2D time-frequency representation, enabling the extraction of speaker-specific details such as formant frequency trajectory, energy changes and co-articulation.[26] A popular method for integrating temporal data to features is to use first and second-order time derivative referred to as delta and double-delta coefficients [18-19], respectively. Also, MFCCs are often enhanced by appending deltas (13 coefficients) and double deltas (13 coefficients), implying 39 features per

frame. These provide additional information about the dynamics of speech signal and can be evaluated as simple changes from the previous frame or weighted sums of numerous previous and future frames.

*4. Deep features* – DL has evinced its efficacy as a potent technique for gleaning high-level features from low-level data. Deep features [24, 27] are extracted from the hidden layer output of a DNN trained on a vast amount of speech data. More precisely, when traditional features (like PLPs, MFCCs, etc.) are passed through the deep models (like deep Restricted Boltzmann Machines, speaker-discriminant DNN, etc.) and reach at a particular hidden layer, the resultant feature is known as deep features, regardless of whether its facet is reduced or retained. The importance of deep features is heavily reliant on the layer from which they are generated. Table II presents a comparative summary of various feature extraction techniques.

## 4. Classification

The other stage is the classification or pattern matching technique or most often known as speaker modelling. For SR, speaker models are created that uses speaker-specific feature vectors [31]. It is trained and saved in the system database using feature vectors, extracted from the speaker's speech sample. The model in text-dependent SR incorporates the temporal relationship between feature vectors. It is dependent on the utterance. Choosing a model depends heavily on the technology, performance requirements, development stage, and computational considerations and training ease. There is primary five ways to categorise the SR model's [8] methodologies as:

### *A. Template model*

This model compares training and testing of feature vectors directly. The degree of deviation [32] denotes the level of resemblance. Vector quantization and Dynamic Time Warping (DTW) are two examples of aforesaid model for text-dependent and text-independent SR.

### *B. Stochastic model*

In this model [8], every speaker is described as the probabilistic source with a defined but variable probability density function. Matching is achieved by comparing the likelihood of the test utterance based on the model.

### *C. Generative model*

During the training phase, statistical characteristics of a speaker-specific speech signals are acquired. The training dataset is used to create a joint probability distribution, then utilised to forecast future output. The most prevalent models based on this approach are HMM and GM

*1. Hidden Markov Models (HMM)* – Recent SR systems use probabilistic / stochastic modelling methods, such as Hidden Markov model [4, 6]. The probabilistic process has optimized the recognition accuracy. In a stochastic
model, a pattern is matched by computing the likelihood of the feature vector for a given speaker model. It models the time variations of the spectral features of speech signal. The parameters of the model can be generated automatically through trained data. It is a finite state machine which has the following specifications - a set of hidden states, a result that is visible from the state, transition probability between the states, discharge probability and preliminary state probability. It is made up of a series of changes that occur between various stages. Its transition is permitted to the next correct state or to remain unchanged. During the training phase, speech signal is used to produce the HMM parameters and for identification, the likelihood of the acquired feature sequence is then calculated with reference to the speaker HMMs.

*2. Gaussian Mixture Model (GMM)* – In 1995 Reynold proposed the GMM [9] for SR. It is an extended version of VQ model. The GMM technique is a type of generative model. It employs an expectation maximization (EM) algorithm [6] which moves the initially random features of the distributions closer to those of observed data iteratively. It accepts a sequence of vectors from the extracted feature vectors (as input) and create one model per speaker. A GMM simulates the speaker using a combination of Gaussian probability density functions (PDFs) computed using the feature coefficients. GMM denoted by $\lambda$, is the PDF which represents a weighted sum of Gaussian parameter density. For the $D$-dimensional feature vector designated by $x$, the mixture density is expressed as a weighted sum of $M$ components, the Gaussian density is shown below:

$$p(x/\lambda)=\sum^{M}_{(i=1)} w_i p_i (x/\mu_i ,\sum i ), \quad (6)$$

Where $p_i (x/\mu_i , \sum_i)$ is the component density, $w_i$ is the weight and M is the number of Gaussian components.

*3. Gaussian Mixture Model-Universal Background Model (GMM-UBM)* – GMM needs large amount of data to model the speaker, therefore to overcome this drawback GMM-UBM was introduced by Reynolds et al. for the SR task. In UBM-GMM system, at first data collected from the registered speaker is grouped and then training of UBM is done [9]. It behaves as the speaker-independent system. The speaker-dependent system is then created using the UBM by using maximum a posteriori (MAP) adaptation technique from speaker-specific trained speech signal. The advantage of the UBM oriented model technique is that it boots performance for small amount of speaker-dependent data.

*D. Discriminative model*

Data pertaining to the target and impostor speakers are engaged in the training phase, and it estimates the parameters that identify the features of the target speaker from the features of the imposter speaker. Also, this model quantifies a parametric model from the input

**Table II.** Feature Extraction Techniques

| Techniques | Property | Disadvantage |
|---|---|---|
| MFCC [18, 19] | It is fast and extracts vital information from the signal.<br>It is non-linear and closely matches the human auditory perception system. Thus, gives better speaker discrimination. | It cannot map continuous speakers effectively.<br>Not resilient to noise. |
| ΔMFCC[18, 19] | Add dynamic information to the static cepstral features. Improving ASR recognition accuracy. | It is not robust to noise and reverberation. |
| ΔΔ MFCC [18,19] | It is very simple to calculate and provide often a clear benefit over the instantaneous features.<br>Improve speaker verification. | It is not robust to noise |
| Bark frequency cepstral coefficient (BFCC) [28] | The Bark scale offers flexibility to the Mel scale in terms of perceptual motivation. | Due to high rate of erroneous rejection, this method is unsuitable for high security application. |
| GFCC [20] | High robustness against acoustic change and noise.<br>Better resolution at low frequencies | It is not a scale invariant feature. |
| LPC [6] | Reflects the vocal tract and a precise, reputable means of getting features.<br>Robust and extract features from low bit rate audio signals. | It cannot discriminate among words, having a kin sound speaker.<br>Feature coefficients have a high degree of correlation.<br>Due to LPC's assumption that the provided signal is stationary, it may be unable to effectively assess local occurrences. |
| LPCC [1, 11] | Cepstral analysis eliminates data with significant similarity.<br>Compared to LPC, this method is more reliable. | It cannot adequately depict speech because it presumes the provided signal is static and cannot effectively evaluate event.<br>During the testing phase, no prior information can be retained. |
| PLP | Disparity between voiced and unvoiced inputs is lessened.<br>Vocal tract length is unaffected.<br>Resulting feature vector has fewer dimensions than the original. | Spectral balance of the formant amplitudes has a significant impact on the feature vector generated.<br>Spectral balance can be easily tipped by factors such as channel, noise, and the equipment used to obtain the input signal. |
| RASTA-PLP | It is resilient and eliminates sluggish and quick changes in the speech signal. | It does not function well for noise-free speech samples. |

| | | |
|---|---|---|
| | Records low modulation frequencies that correlate to speech in a signal. | |
| DWT [1] | Consider signal's temporal as well as frequency information.<br>Successfully executes de-noising operations<br>Provided signal is reconstructed accurately from the decomposed elements. | Using same base wavelet for all input signal, makes the technique rigid. |
| Maximum-Likelihood Linear Regression (MLLR) | It decimates the flaws of spectral mapping and model mapping techniques.<br>Facilitate speaker normalization.<br>On the premise of this likelihood, estimators are generated. | One of the key concerns with MLLR adaptation is estimating the regression coefficients with confidence based on the available dataset. |
| Feature Space MLLR (fMLLR) | Pivotal acoustic feature for DNN/HMM hybrid SR models.<br>A constrained MLLR, include an extensive transformation matrix and observation vector. | For restricted adaptation data, the transformation vectors have a propensity to overfit the data source. |
| Empirical Mode Decomposition (EMD) [29] | Simpler to grasp and apply than other spectrum approaches.<br>Resistant to non-linear and non-stationary input. | It is noise-sensitive and rigorous mathematical framework is lacking. |
| Power Normalized Cepstral Coefficients (PNCC) [30] | Power-law nonlinearity substitute traditional log nonlinearity used in MFCC.<br>Recognition accuracy is better than MFCC, RASTA-PLP and PLP even in noisy environment.<br>Reliant on input speech's explicit envelop estimation. | Computational cost is greater than MFCC. |

training set and its output vectors. Examples are ANN and SVM.

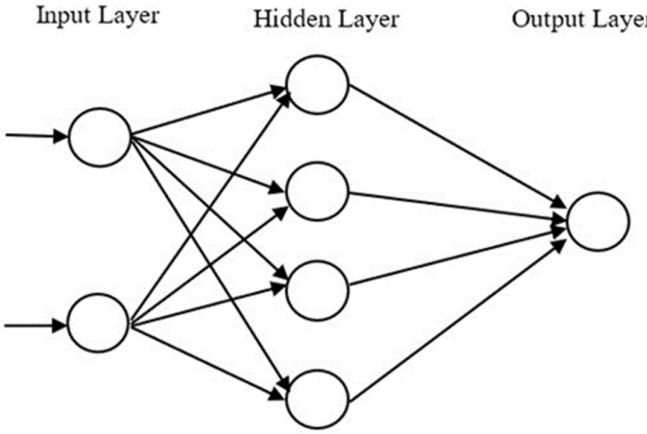
*1. Support Vector Machine (SVM)* – SVM [33] was developed by Vapinik and is a statistical learning theory-based technique. In recent years, SVMs were extensively preferred to resolve binary classification problems. It produces a hyperplane in a multidimensional vector space that is used to split vectors belonging to two distinct classes. SVM seeks the hyperplane with the greatest margin, often known as the maximum margin hyperplane. It is a robust discriminative classifier8, used for prosodic, spectral and high-level features. It has shown robust performance in speaker identification. When used in combination with GMM the accuracy increases. The increased performance is obtained with SVM due to its ability to classify unseen data. The authentic trained data is mapped to a high dimension using nonlinear mapping. The mapping of the input space to the higher dimensional feature space is accomplished using kernel functions. It seeks the linear optimal separation hyperplane, also known as the decision boundary, to split data from two classes. SVM use support vectors to locate this hyperplane. During testing, a classification score is obtained by computing the distance of the test sample in relation to the hyper-plane.

E. *Hybrid model*

This is a blend of generative and discriminative models. Additionally, this model can also be utilised for classification purposes, combining the advantages of both types. One example is HMM with ANN. The hybrid combination of DL and machine learning (ML) can help to decimate the over-fitting problem often faced due to less data. Table III presents a comparative summary of various traditional, ML and DL classification techniques [1, 8, 11].

*F. Neural Network and Deep Learning*

Neural network is based on the human brain and are designed to function similarly to the way in which biological neurons communicate with one another. As depicted in the Fig. 4, it consists of a node layer with an input layer, one or more hidden layers, and an output layer. Every node in the network has a weight and a threshold associated with it. Any node whose output is greater than the threshold value will be activated and will communicate its data to the next tier of the network. If not, no data passes to the next tier.




The accuracy of these networks is enhanced through exposure to training data. However, after being optimised for precision, these learning algorithms become potent tools in artificial intelligence and computer science, enabling rapid data classification and clustering. DNN were initially applied to image recognition and their initial applications to SR were frequently limited to a single chunk of the pipeline, either front or back-end. Subsequently, SR methods became wholly DNN based, so-called end-to-end approaches. The words DNN and DL are frequently used interchangeably, but it does not always entail the other. It is widely accepted that a neural network essentially has two or more hidden layers in order to be considered "deep." DL (while it can be implemented to non-NN approaches, referred to as unconventional DL), can be thought of as the collection of techniques, algorithms and approaches which makes DNN training initially possible and then widespread.

*1.* Deep Belief Network (DBN) – It was first described in [34], as generative models consisting of numerous interrelated secluded hidden layers, in which units from one layer are connected only to units from the preceding and subsequent layers, without any lateral connections. Weights of these connections are optimized using greedy layer-by-layer training. The first layer serves as the (perceptible) training set, while the first hidden layer feeds its output to the subsequent layer, resulting in the rapid unsupervised training. The hidden units are the stochastic latent parameters with an output that is a weighted sum of the inputs that have been transferred through an activation function, which is commonly sigmoid or rectified linear.

*2.* Deep Auto-encoders (DAE) - Auto-encoders (AEs) work on the principle of mapping input patterns upon themselves by moving through a simplified parameter representation. It is accomplished by possessing one hidden bottleneck layer that is narrower than the previous ones. DAE is also known as auto-encoder stacks, are made up of multiple stacked AE networks, with the bottleneck layer of one AE functioning as the input layer for the succeeding (narrower) encoding layer, accompanied by a symmetrical decoding framework. They are trained in a manner similar to DBNs, with an unsupervised pre-training pass accompanied by supervised tuning through back-propagation [35]. SR systems are affected by environmental factors like speaker distance from the input device, noise, reverberation, and so on. AEs are used to remove environmental influences from speech data [7].

*3.* Convolutional Neural Networks (CNNs) – CNN [11] is a feed-forward networks which extract successively higher level features from the given input data. A prevalent CNN architecture consists of substitute pooling and convolutional layers, followed by fully connected layer. The convolutional function is expressed mathematically using the following representation: $g(x, y) = i(x, y) * h(x, y)$ (7)

Here i(x, y) denotes the input signal, h(x, y) denotes the applied filter and g(x, y) denotes the resulting convolved filter. On all of the receptive fields, the same filter with

**Table III.** Classification Techniques

| Classification Techniques | Property | Disadvantage |
|---|---|---|
| HMM [4, 6] | Models the speech signal's time distribution.<br>It is incredibly simple to put into action.<br>It is capable of processing inputs of varying lengths.<br>Discrete and continuous sequences are modelled | It ignores long term dependencies since the current state is presumed to be independent of any previous state. |
| ANN [11] | It is durable; self-organization and learning are key features of this algorithm.<br>It is flexible and adaptable to changing conditions. | If it overtrains, it may become mired in local minima issues. |
| RBF [11] | Simple to put into action and durable.<br>It is capable of distinguishing between various terms in a vocabulary with ease. | Shifting in time has no effect on it. |
| RNN [11] | It can easily handle varied length inputs.<br>Capable of retaining temporal dependencies.<br>Weights can be shared between different time steps. | Computationally costly and takes a long time to train.<br>More susceptible to disappearing and exploding gradient. |
| LSTM [22] | Type of RNN, flexible enough to cope with long-term dependency.<br>Deal with the RNN's vanishing gradient issue.<br>Permits the preservation of data. | Susceptible to overfitting.<br>Require large memory bandwidth.<br>Impacted by varying initializations of random weights. |
| LSTM-RNN [22] | Accomplish greater recognition rate in acoustic modelling than DNN. | Highly susceptible to static inputs. |
| CNN [11] | It is extremely efficient in identifying essential features of a signal.<br>Requires minimal training time. | It cannot capture temporal features properly.<br>It can't deal with variable size input. |
| GMM [9] | Maximum log likelihoods are used to make recognition judgment by examining frame attributes associated with a given speaker model. | It needs adequate data to accurately simulate the speaker. |
| SVM [33] | It is quite durable.<br>Not prone to over-training or being stuck in local minima<br>Can be used with high-dimensional vector input. | It can only accept fixed length inputs.<br>Processing expenses rise as the number of output classes increases |
| VQ | It is very efficient in data compression.<br>For user identification, the distance between average of the speech data sample and training data sample is evaluated. The user identity is determined by the smallest distant set. | Limited size of dataset. |

| | | |
|---|---|---|
| UBM | It is a speaker-independent GMM which depicts the speech features.<br>Used in speaker verification system and fast to train and test. | Alleviate the issue of inadequate training data and unobserved data. |
| UBM-GMM | Statistics are influenced by the amount of mixtures and the quantity of vectors to score.<br>Probabilistic in nature.<br>Powerful tool for recognising speakers. | Variation in microphone or acoustic surroundings affect results between training and recognition data. |
| i-vector (identity vector) [36] | Reduce high dimensional space to low dimensional space for speaker and channel variations by simple factor analysis.<br>Speaker and channel share the same area. | Efficiency of the system degrades for brief utterance.<br>Susceptible to language and channel incompatibility |
| x-vector [36] | Built on DNN embeddings in which neural networks are programmed to differentiate among speakers.<br>Speaker embedded at the segment level. | Lack robustness to intra-speaker variations. |
| d-vector (deep vector) [11] | Collect acoustic characteristics from the raw data.<br>Feed forward propagation is used to obtain the output activation of each frame from the previous hidden layer. | Speaker data persist in long-term chunks; therefore, training frame wise doesn't work well. |

the same set of weights is used. This property enables CNN to acquire the majority of the features of the speech
signal without resorting to a vast number of weights.

*4. Long short-term Memory Recurrent Neural Networks (LSTM-RNN)* – RNN [11] is a type of neural network architecture that follows a cyclical path. RNNs are intrinsically able to retain historic information, thus they are suitable for tasks involving the modelling or classification of dynamic or time-varying data. RNN is a preferred DL technique for SI because it frames signals in a short time frame for feature extraction. Moreover, training RNN is a difficult task due to vanishing gradients, which hampers the overall results. As a consequence, to decimate the above drawbacks, Long short-term memory (LSTM) [22] was proposed.

LSTMs can learn long-term dependencies and are not affected by vanishing or exploding gradients. They also do not require unsupervised pre-training. It is accomplished via memory units that contain constant error carousels with error derivatives of 1.0 which can neither be exploding nor disappearing. Few non-linear units, known as forget, keep, update, and output gates, are in charge of retaining or expelling the present and previous information. LSTMs are capable of accounting for events that took place thousands of time steps ago. LSTMs underpin the majority of successful RNN approaches.

## 5. Programming Toolkits

Over the years, the renaissance of neural network research has necessitated the development of versatile and efficient DL frameworks [7]. As a corollary, a slew of open-source implementations [7, 11] are actively being incorporated. TensorFlow [37], PyTorch [38], ASVtorch [39], Alize [40], Kaldi [16] and HTK [41] are a few of the most prominent options. Using these toolkits will benefit DL research and production because they allow for flexible model creation at a big scale and in varied situations. The support for programmable network structure considerably helps develop and test a study proposal, adding to the boom of DL research.

However, in the realm of SR, an open-source toolkit called Kaldi has gained much traction [16, 42, 43]. It includes a collection of useful tools developed in C++ that facilitate the development of efficient ASR systems along with speaker identification. It is primarily used to implement i-vector and x-vector system. One significant advantage of Kaldi over other voice recognition toolkits is the availability of numerous state-of-the-art solutions for diverse datasets, which substantially decreases the time required to begin a project. It provides three distinct neural network-based model development and training implementations. Among these three, the "nnet3" is meant to provide programmable network structures, i.e., model construction [15] without explicit coding. Few more DL toolkits are discussed in Table IV. It is indeed evident that DL's arrival has sparked numerous new applications for ML and AI in general [13]. DL has made it possible to break down tasks in the simplest ways to aid machines most effectively. A DL framework

based on Python, such as Kaldi is state-of-the-art. Before selecting the most acceptable DL framework, users should examine speed, resource consumption, utilisation and the trained model's consistency.

## 6. Blockchain with Speaker Recognition

Blockchain [46, 47] has the potential to offer SR with advantageous characteristics, including immutability, accountability, availability, and universal access. The utilisation of blockchain technology offers significant advantages as shown in Fig. 5, particularly in the field of biometrics, by providing the means to safeguard biometric templates and ensure anonymity within SR systems.

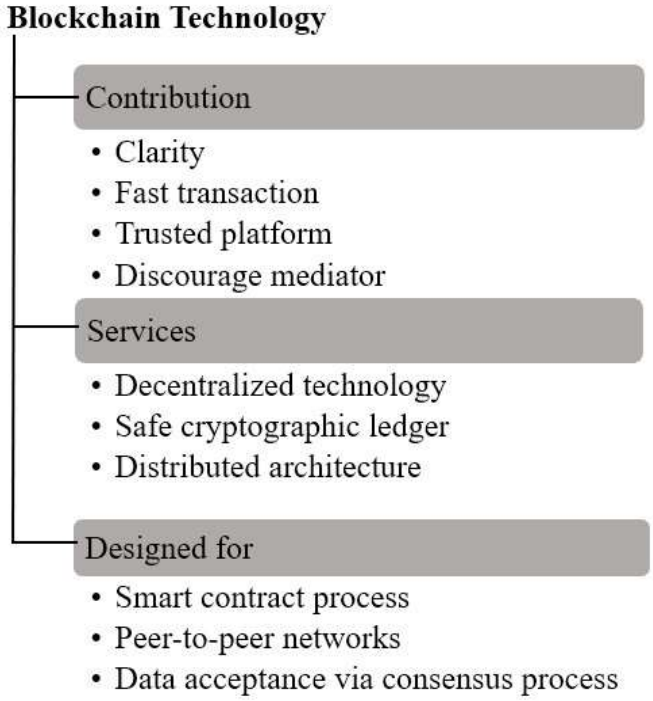



**Figure 5**. Advantage of Blockchain Technology

Furthermore, this technology has the capability to utilise speaker recognition in other ways, such as enhancing existing decentralised digital identification systems on blockchain. Another intriguing use pertains to intelligent devices. A smart device refers to any electronic or physical item that has the capability to access a blockchain and execute actions and choices depending on the data stored within it. Implementing a biometric-based authentication [51] methodology has the ability to greatly enhance the existing degree of security.

Recently, Lee et al., [48] has proposed Ethereum Blockchain-based Distributed Biometric Authentication System, (BDAS). By dividing a biometric template into fragments and maintaining them using a blockchain method, BDAS increases the security and reliability of existing biometric authentication systems. Specifically, it improves biometric information security via blockchain-based distributed management, increases authentication reliability through blockchain-based decentralised authentication, and ensures transparency of biometric information flow through a blockchain-based audit mechanism. The examination of BDAS shows that, when compared to existing approaches, it delivers secure and dependable authentication with low performance impact in real-world circumstances.

Similarly, Upadhyay et al., [49] presented speaker recognition using MFCC feature extraction and Blockchain for data security. The suggested method was examined using several feature removal techniques in order to extract the best value in error rate and accuracy. The proposed strategy improves the rate of correct evidence collecting while reducing loss and authentication problems.

**Table IV.** Programming Toolkits

| DL Toolkit | Programming Language | Property |
|---|---|---|
| YAAFE [11] | Python | Simplicity is achieved by correctly exploiting redundancy in |

| | | |
|---|---|---|
| | | the feature evaluation.<br>Convenient configuration with parametrizable options for all features. |
| LibXtract [11] | C | Consists of an extensive corpus of audio characteristics.<br>Makes use of three different types of audio characteristics: vector (which returns an array of audio features), scalar (which returns a single value), and delta (which returns a temporal dimension). |
| HTK [42] | C | HMM based toolkit, used for SR model |
| Tensorflow [37] | C++,<br>Python | Employed in DL application to quickly adopt modern algorithms while maintaining the same server architecture and APIs in place |
| PyTorch [38] | Python | Employed in DL application as it has a better user interface and is simpler to operate.<br>Enables data to be exchanged with outside repositories |
| Kaldi [16, 43] | C++ | Acoustic modelling and feature extraction. |
| PyTorch-Kaldi [44] | Python | PyTorch and Kaldi libraries are combined. |
| Alize [16] | JAVA | Includes all functions required to use Gaussian mixtures. |
| ASVtorch [39] | Python | Use PyTorch machine learning approach. |
| Keras [3] | Python | Lightweight, user-friendly and simple in design.<br>Common application is in the field of classification, text generation, tagging, summarization, translation and SR. |
| CNTK [45] | Python,<br>C++ | Efficient for SR and faster than Caffe and TensorFlow in training DL models.<br>DL architectures like CNN, RNN, and LSTM are implemented using this toolkit. |
| Caffe [46] | Python,<br>C++ | Notable for speed and defined by configuration rather than hard coding.<br>Setup and training are simplified as no additional network infrastructure is required. |
| MXNet [46] | Python,<br>C++, Julia | It is an effective and versatile library for DL.<br>It enables the developers to take advantage of the GPU. |

Furthermore, Zhang et al., [50] suggested a methodology for protecting black-box voiceprint recognition models that incorporates active and passive protection. It embeds important information within the Mel spectrogram to produce difficult-to-detect and delete trigger samples, which it injects into the host model as a watermark. As a result, the voiceprint recognition model's copyright protection performance improves. To prevent unauthorised users from using the model, the model's index number and encrypted model data are placed on the blockchain, and then an exclusive smart contract is constructed for limiting access to the model. The results of the experiments reveal that this framework efficiently protects the copyrights of voiceprint recognition models and prohibits unauthorised access.

## 7. Challenges of speaker recognition system

The reliability of SR system is evaluated on range of well-known factors [6, 7, 11, 24].

- Speaker, pronunciation, geography, speech rate, context, channel, and environmental variation are the most notable. When constructing SR systems and effective models must be built to achieve adequate identification/verification accuracy irrespective of these variances.
- It is usual to see variances between classes and within classes in the inter speaker and intra speaker variations in pattern categorisation tasks.
- Variations in the recording medium and surroundings cause channel variation.
- Multi-speaker audio can be tricky due to overlap speech. With this kind of cross-talk, it's nearly impossible to tell which speaker is speaking at any specified instant.

- Although text plays a minor role, compared to text-dependent systems, text-independent systems are arduous to develop.
- Since the number of speakers in the database grows, we need to concede SR, which substantially affects the system's performance.
- Higher-level SR system design includes automatic word lexicon generating algorithms or processes, automated speaker segments algorithms, excellent utterance verification/rejection algorithms and human performance in SR activities.
- As speech is a behavioural biometric utilised for SR, it signifies that the speech features may change with age, sickness, exhaustion, tension, and so on.
- Traditional approaches to SR rely on spectrum-related features based on very small periods of time slices of speech. Approaches rely on such knowledge, lack robustness to channel imbalance and fail to grasp longer-term characteristics of how an individual speaks, such as the speaker's word structure and patterns in speech prosody (the timing, pausing, and intonation of speech).
- Significant advancements in SR technology, short speech, background noise, channel distortion, and other concerns continue to significantly impact SR performance in real-world applications.
- Construct Long-range features, which occur less often than relatively short-range characteristics, can give significant supplementary information for 30-second, training and test speech surges.
- Spoofing with audio recording is a critical drawback when trying to use SR. This can be prevented by asking the system to repeat a randomly produced phrase. This makes impossible for an impersonator to predict the randomized word and thus perform a replay spoofing attack.

## 8. Conclusion

This paper analyses the concepts and advancements made for the human voice for speaker recognition. The key benefit of speech-based biometrics is that it accomplishes recognition without explicit visual or physical interaction with the individual. To design a robust and efficient SR system, extraction of efficacious speech feature is necessary to increase the accuracy of SR. The problem emerges because the speech features changes with time and surroundings, which compel it more challenging to recognise precisely. The variation or mismatch in biometric (audio) data at training and testing stage, caused due to age, sickness, emotion etc. should be minimized. Therefore, this paper outlines developmental history, technological progress, concepts to explore state-of-the-art techniques for feature extraction, robust speech features, traditional and neural network based classification techniques and SR toolkits. Despite the vast amount of work done in SR area, it was found that very less research has been done for integration of blockchain with speaker recognition, especially for people with disabilities. Thus, more research should be focused to design a robust, convenient and reliable recognition system for people with disabilities which will significantly perk in the advantages of digital era to enhance their well-being and quality of life.

## References

[1] Hanifa, Rafizah Mohd, Khalid Isa, and Shamsul Mohamad (2021) A review on speaker recognition: Technology and challenges. Computers & Electrical Engineering 90: 107005. https://doi.org/10.1016/J.COMPELECENG.2021.107005

[2] B. Homayoon (2011) Speaker Recognition, Springer Science & Business Media.

[3] Furui, Sadaoki (2004) Fifty years of progress in speech and speaker recognition. The Journal of the Acoustical Society of America 116: 2497-2505. https://doi.org/10.1121/1.4784967

[4] Greenberg, Craig S., et al. (2020) Two decades of speaker recognition evaluation at the National Institute of Standards and Technology. Computer Speech & Language 60:101032. https://doi.org/10.1016/J.CSL.2019.101032

[5] Kabir, Muhammad Mohsin, et al. (2021) A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. IEEE Access 9:79236-79263. https://doi.org/10.1109/ACCESS.2021.3084299

[6] Ohi, Abu Quwsar, et al. (2021) Deep Speaker Recognition: Process, Progress, and Challenges. IEEE Access 9:89619-89643. https://doi.org/10.1109/ACCESS.2021.3090109

[7] Kinnunen, Tomi, and Haizhou Li (2010) An overview of text-independent speaker recognition: From features to supervectors. Speech communication 52(1):12-40. https://doi.org/ 10.1016/J.SPECOM.2009.08.009

[8] Al-Ali, Ahmed Kamil Hasan, et al. (2017) Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access 5:15400-15413. https://doi.org/10.1109/ACCESS.2017.2728801

[9] Srivastava, Sumit, Mahesh Chandra, G. Sahoo (2019) Speaker identification and its application in automobile industry for automatic seat adjustment. Microsystem Technologies 25(6):2339-2347. https://doi.org/10.1007/S00542-018-4111-Z

[10] Jahangir, Rashid, et al. (2021) Speaker Identification through Artificial Intelligence Techniques: A comprehensive Review and Research Challenges. Expert Systems with Applications 114591. https://doi.org//10.1016/J.ESWA.2021.114591
[11] Reynolds, Douglas A., Thomas F. Quatieri, Robert B. Dunn (2000) Speaker verification using adapted Gaussian mixture models. Digital signal processing 10:19-41. https://doi.org/ 10.1006/DSPR.1999.0361
[12] Zhao, Xiaojia, DeLiang Wang (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification. In: IEEE international conference on acoustics, speech and signal processing 7204-7208.
[13] Abdalrahman, Roaya Salhalden A., Bülent Bolat, Nihan Kahraman (2018) A cascaded voice biometric system. Procedia computer science 131: 1223-1228. https://doi.org/10.1016/J.PROCS.2018.04.334
[14] Cucu, Horia, et al. (2015) Enhancing ASR systems for under-resourced languages through a novel unsupervised acoustic model training technique. Advances in Electrical and Computer Engineering 15(1)-63-68.
[15] Li, Lantian, et al. (2022) CN-Celeb: multi-genre speaker recognition," Speech Communication 137:77-91.
[16] Gonzalez-Rodriguez, Joaquin (2014) Evaluating automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996-2014) Loquens 1. https://doi.org/ 10.3989/LOQUENS.2014.007
[17] Jokic, Ivan, et al. (2015) Automatic speaker recognition dependency on both the shape of auditory critical bands and speaker discriminative MFCCs. Advances in Electrical and Computer Engineering 15(4):25-33.
[18] Sharma, Garima, Kartikeyan Umapathy, and Sridhar Krishnan (2020) Trends in audio signal feature extraction methods. Applied Acoustics 158:1-13. https://doi.org/ 10.1016/J.APACOUST.2019.107020
[19] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye (2006) A fast learning algorithm for deep belief nets. Neural computation 18(7):1527-1554.
[20] Hourri, Soufiane, Jamal Kharroubi (2020) A deep learning approach for speaker recognition," International Journal of Speech Technology 23(1):123-131. https://doi.org/10.1007/S10772-019-09665-Y
[21] Li, Lantian, et al. (2022) A principle solution for enroll-test mismatch in speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 443-455.
[22] Baig, Faisal, Saira Beg, and Muhammad Fahad Khan (2018) Speaker recognition based appliances remote control for severely disabled, low vision and old aged persons. INAE Letters 3(1):1-9. https://doi.org/10.1007/S41403-017-0032-X
[23] Wang, DeLiang, and Guy J. Brown (2006) Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press. https://doi.org/10.1109/TASLP.2023.3244507
[24] Govindan, Sumithra Manimegalai, Prakash Duraisamy, and Xiaohui Yuan (2014) Adaptive wavelet shrinkage for noise robust speaker recognition. Digital Signal Processing 33:180-190. https://doi.org/10.1016/J.DSP.2014.06.007
[25] Wu, Jian-Da, Yi-Jang Tsai (2011) Speaker identification system using empirical mode decomposition and an artificial neural network. Expert Systems with Applications 38(5):6112-6117. https://doi.org/10.1016/J.ESWA.2010.11.013
[26] Tirumala, Sreenivas Sremath, et al. (2017) Speaker identification features extraction methods: A systematic review. Expert Systems with Applications 90:250-271. https://doi.org/ 10.1016/J.ESWA.2017.08.015
[27] Campbell, William M., Douglas E. Sturim, and Douglas A. Reynolds (2006) Support vector machines using GMM supervectors for speaker verification. IEEE signal processing letters 13(5):308-311. https://doi.org/10.1109/LSP.2006.870086
[28] Schmidhuber, Jürgen (2015) Deep learning in neural networks: An overview. Neural networks 61:85-117. https://doi.org/10.1016/J.NEUNET.2014.09.003
[29] K. Jha, A. Jain and S. Srivastava (2023) An Efficient Speaker Identification Approach for Biometric Access Control System. 5th International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, India, pp. 1-5. https://doi.org/10.1109/RAIT57693.2023.10127101.
[30] Jain, Anil K., Robert P. W. Duin, Jianchang Mao (2020) Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence 22(1):4-37. https://doi.org/10.1109/34.824819
[31] Sumit Srivastava, G. Sahoo, Naman Ladha, Mahesh Chandra (2018) A Review on User Identification using Voice as a Biometric Feature. International Journal of Computer Application, USA 11-14.
[32] Hebert, Matthieu (2008) Text-dependent speaker recognition," Springer handbook of speech processing. Springer, Berlin, Heidelberg, pp. 743-762.
[33] Larcher, Anthony, et al. (2014) Text-dependent speaker verification: Classifiers, databases and RSR2015: Speech Communication 60:56-77 . https://doi.org/10.1016/J.SPECOM.2014.03 .001
[34] Devi, Kharibam Jilenkumari, and Khelchandra Thongam (2023) Automatic speaker recognition from speech signal using bidirectional long-short-term memory recurrent neural network. Computational Intelligence 39(2):170-93. https://doi.org/10.1111/coin.12278
[35] Sahoo, Tushar Ranjan, and Sabyasachi Patra (2014) Silence removal and endpoint detection of speech signal for text independent speaker identification. International Journal of Image, Graphics and Signal Processing 6:27-35. https://doi.org/10.5815/IJIGSP.2014.06.04

[36] Farsiani, Shabnam, Habib Izadkhah, and Shahriar Lotfi (2022) An optimum end-to-end text-independent speaker identification system using convolutional neural network. Computers and Electrical Engineering 100:107882. https://doi.org/10.1016/J.COMPELECENG.2022.107882
[37] Gan, Zhen-ye, Yue Yu, and Min Luo (2022) A tibetan-dependent speaker recognition method based on deep learning. Multimedia Tools and Applications 81:30821–30840. https://doi.org/10.1007/S11042-022-12540-9
[38] Lee, Kong Aik, Ville Vestman, and Tomi Kinnunen (2021) ASVtorch toolkit: Speaker verification with deep neural networks. SoftwareX 14:100697. https://doi.org/ 10.1016/J.SOFTX.2021.100697
[39] Sreehari, V. R., and Leena Mary (2022) Automatic short utterance speaker recognition using stationary wavelet coefficients of pitch synchronised LP residual. International Journal of Speech Technology 25(1):147-161. https://doi.org/10.1007/S10772-021-09895-Z
[40] Li, Dongdong, et al. (2022) TRSD: A Time-Varying and Region-Changed Speech Database for Speaker Recognition. Circuits, Systems, and Signal Processing 41(7):3931-3956. https://doi.org/10.1007/S00034-022-01964-1
[41] Zhang, Xingyu, et al. (2023) Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition. Complex & Intelligent Systems 9(1):65-79. https://doi.org/10.1007/S40747-022-00782-X
[42] Zhao, Hong, et al. (2022) Research on x-vector speaker recognition algorithm based on Kaldi. International Journal of Computing Science and Mathematics 15(3):199-212. https://doi.org/ 10.1504/IJCSM.2022.124725
[43] Paszke, Adam, et al. (2017) Automatic differentiation in pytorch. In 31st Conference on Neural Information Processing Systems, USA.
[44] Pawar, Rupali V., Rajesh M. Jalnekar, and Janardan S. Chitode (2018) Review of various stages in speaker recognition system, performance measures and recognition toolkits. Analog Integrated Circuits and Signal Processing 94(2):247-257. https://doi.org/10.1007/S10470-017-1069-1
[45] Sarkar, Achintya Kumar, et al. (2019) Time-contrastive learning based deep bottleneck features for text-dependent speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(8):1267-1279. https://doi.org/10.1109/TASLP.2019.2915322
[46] Delgado-Mohatar, Oscar et al. (2020) Blockchain and biometrics: A first look into opportunities and challenges. In Blockchain and Applications: International Congress, pp. 169-177. Springer International Publishing.
[47] Jang, Hyeji, and Sung H. Han (2022) User experience framework for understanding user experience in blockchain services. International Journal of Human-Computer Studies 158:102733.
[48] Lee, Youn Kyu, and Jongwook Jeong (2021) Securing biometric authentication system using blockchain. ICT Express 7(3):322-326. https://doi.org/10.1016/J.ICTE.2021.08.003
[49] Upadhyay, Shrikant et al. (2022) Feature Extraction Approach for Speaker Verification to Support Healthcare System Using Blockchain Security for Data Privacy. Computational and Mathematical Methods in Medicine Article ID 8717263, 12 pages. https://doi.org/10.1155/2022/8717263
[50] Zhang, Jing, Long Dai, Liaoran Xu, Jixin Ma, and Xiaoyi Zhou (2023) Black-Box watermarking and blockchain for IP protection of voiceprint recognition model. Electronics 12:3697.
[51] K. Jha, S. Srivastava and A. Jain (2024) Cryptanalysis of a Biometric based Anonymous Authentication Approach for IoT Environment. International Journal of Microsystems and IoT, 2(2), 591–597. https://doi.org/10.5281/zenodo.10804461