Text Independent Biometric Speaker Recognition System and its future Scope

Purushottam Rath

Department of Computer Science and Engineering

Sharda University

Abstract

The process of recognizing a person by their voice is known as Automated Speaker Recognition (ASR). Biometric security technology includes speaker recognition. The subject of biometrics is individuality or human traits. The use of a person's voice as a biometric verification or realistic authentication is used to identify that individual. The term "voice biometrics" refers to measurements of a person's behaviour or physiological state. The use of physiological biometrics, which include the iris, face, fingerprints, retina, ear, and DNA, contrasts with the use of voice, signatures, keystrokes, typing, and other behavioral biometrics. Vocal biometrics is a new area of research nowadays. As well as in person, humans may identify a speaker by hearing their voice on the phone or another digital device.

This fundamental human ability has been utilised in the development of automatic speaker recognition (ASR), a speech biometric authentication technique. An ASR can determine who is speaking by analysing speech signals and characteristics derived from speaker sounds. ASR is being considered as an exciting study area because to its importance to voice biometrics. ASR is used for identify people from their voice. It is a way to repeatedly identifying a speaker who is speaking by the information that is counted in waves. ASR methods makes it possible to use the speaker’s speech to verify the identity of the individuals. It has many applications like voice dialling, remotely accessible computers, banking by telephone, information services, etc.

The individuals can identify a speaker by hearing their voice, over a telephone or any kind of electronic devices. An ASR recognizes speakers by analysing speech signals and characteristics extracted from speaker’s voices. ASR has recently become an effective researcharea as an essential aspect of voice biometrics. Specifically, this literature survey gives a briefintroduction of ASR and provides an overview of the general architectures dealing with speakerrecognition technologies, and supports the past, present, and future research trends in this area.This paper briefly describes all the main aspects of ASR, such as SI, SV, SD etc.

Biometric verification is used to recognize an individual through his/her voice individual characteristic. Voice biometric includes behavioural or physiological measurements of individual. Behavioural biometric is performed by Voice, Signature, Keystrokes, and Typing etc. whereas physiological biometric includes iris, face, retina, fingerprints, ear, DNA etc. Now a days voice biometric is emerging research area.

Further, the performance of current speaker recognition systems is investigated in this survey with the limitations and possible ways of improvement. This study provides a brief summary of ASR and the basic principles underpinning speaker recognition technology while supporting past, present, and future research trends in this area.

Keywords: ASR, DNA, Retina, Ear, Fingerprints

Introduction

Biometric analysis is a branch of science that leads with the statistical analysis of any biological data of an individual human being [1]. Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on [2]. Automatic speech recognition has greatly contributed to the development of artificial intelligence, which seeks to create very flexible methods of handling the machine, this allows the user to communicate and exchange information without using known input/output modules such as the keyboard. Voice-based input/output techniques are very useful in several areas, such as the care of disabled people, the use of cars, in particular when driving, distress calls, etc [3].

Automatic Speaker Recognition (ASR) is a biometric system that is used to authenticate the uniqueness of the user, by analyzing the specific characteristics produced from user's speech. These systems are used in mobile phones, Alexa, and Google Home to authenticate the voice of the individual for activating various inbuilt applications viz banking, call centres, playing music etc. Keeping in view the wide applications of ASR, its security becomes a very vital and necessary aspect. Voice biometrics, particularly for the purpose of voice authentication, is an emerging field in terms of security. These biological traits have distinguishing and acceptable vocal qualities by which ASR is able to identify a person by his/her voice; even despite any changes in the surrounding affecting the input voice. As the method used in ASR is voice-based biometric which is very practical, affordable, and accurate, it is widely used and accessible in the modern era. In terms of the security, voice biometric speaker recognition needs more improvement so that the attacker is unable to spoof the original voice. An important application of speaker recognition technology is "forensics" [4]. A term "forensic science" is used for the field of examining and gathering information about the attacker. It is used for the legal verification of the user. "Computer forensics" is also known as "digital cyber forensics." It is a branch of "digital forensic science." The aim of digital forensics is to recognize, examine, and recover specific information. Digital forensics is a new area for crime investigation. Voice identification is done using "forensic phonetics," also known as "forensic linguistics." It is carried out based on the voice's audio characteristics. Realistic speaker recognition is used for locating a speaker using forensic linguistics. This duty is crucial and difficult. A speaker recognition application is forensic speaker recognition.

In voice recognition, two parameters are always considered, i.e., speaker verification and speaker identification. Voice recognition refers to the process in which a person's authentication is done with the help of his own voice, also known as "Biometric Identification Technique" (BIT). Human's biological traits are used in this technique to identify and verify a person. Authentication is the process of confirming a user's identity when they request Physical Access (PA) or Logical Access (LA) into the ASR.

1.1 Major contribution of the work

In SV, the appealed uniqueness is matched against specific speaker's voice model while in speaker identification system tries to match an unknown speaker against the entire voice database. Speaker recognition systems can be divided into text-dependent and text-independent ones. In text-dependent systems, suited for cooperative users, the recognition phrases are fixed, or known beforehand. For instance, the user can be prompted to read a randomly selected sequence of numbers. In text-independent systems, there are no constraints on the words which the speakers are allowed to use [5]. The human speech signal contains different types of information that makes for authentication of that human voice. A speech signal uttered by a person can identify person. By using a speech signal mainly three kinds of recognition are performed; speech recognition (what is spoken), speaker recognition (who is speaking) and language identification (identifying the speaker's spoken language). An ASR is likely to face a challenge named as "Phonetic Variability". In general, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition. Changes in the acoustic environment and technical factors (transducer, channel), as well as "within-speaker" variation of the speaker him/herself (state of health, mood, aging) represent other undesirable factors. In general, any variation between two recordings of the same speaker is known as session variability. Session variability is often described as mismatched training and test conditions, and it remains to be the most challenging problem in speaker recognition [6].

This research paper comprises of the following sections viz Background Knowledge, Literature survey, Proposed Methodology, Classification of Speaker Recognition (SR) system, Performance Evaluation of Voice Biometric System, Problems with ASR systems, Results and Discussion and finally conclusion.

2. Background Knowledge

Speaker recognition (SR) technology allows for the identification of a person based on their distinctive speech. Each person's vocal tract, larynx, and other voice-producing organs are different; hence no two persons have the same voice. In addition to these physical differences, every person has a distinctive speaking pattern, way of pronouncing words, preferred lexicon, and other traits. Owing to all of these considerations, speech recognition technology can be utilized in addition to fingerprint and retinal scans as a biometric. The SR work is divided into the "Speaker Identification" and "Speaker Verification" portions. The term "Speaker Verification" refers to the process of verifying that the speaker is who they say they are. Due to uneven training and testing models, performance degradation in SR tasks gradually appears. Among the variables influencing performance are background noise, channel effect, and speaker-based variability. A speaker's effectiveness can be influenced by a variety of elements, including health, phonation, vocal effort, and emotional content. Most humans frequently express their emotions while speaking with others, making SR in emotional contexts a significant research problem in human-computer interaction. Hence, even though it can be challenging, it's critical to be able to recognize when someone is speaking from an emotional perspective, such as when they're sad, furious, or happy.

Many issues, like enhancing speech recognition in telecommunications, can be addressed by this research. The call centers where this study will have the biggest impact will see an improvement in customer service thanks to the deployment of emotionally intelligent automated technologies. It is not always possible to speak in neutral mode in real-world settings, even though neutral speech is utilized for training and testing in most SR-related tasks. This study found that when training and testing of emotions were not conducted in unison, SR

systems performed worse. The key factors contributing to this performance decline are the mismatch of some emotions in the test utterances and the speaker models, as well as the articulation patterns of some of those emotions, which produce significant intra-speaker vocal variability.

An emotional normalization technique is used to remove the emotion bias in scoring. In their research, Raju et al. report on SR in emotional contexts using a fused database made up of the IITKGP-SESC: Hindi, IITKGP-SESC: Telugu, and the German Emotional Speech Database (EMO-DB). Analysis of source and system features reveals that the emotional state of the speaker has a stronger influence on the source features. We contrast several modelling techniques. In this study, the role of emotional information in creating the SR system's correctness is examined in order to improve its functionality.

3.RELATED WORK

In the past, many algorithms were developed to authenticate a person. Mostly, these algorithms were used to extract the features from the input audio signal. Below, we shall review the proposed techniques, which were used for the authentication of the person by his/her voice.

One of the first category of the work was proposed by Reynolds and Rose**.** The authors suggested the use of Mel-Frequency Cepstral Coefficients (MFCCs) as features and a Gaussian Mixture Model (GMM) to authenticate the speaker. MFCCs characterize adequately the envelope of the short-time power spectrum of the signal. Despite being sufficiently resilient to noisy conditions, their applicability in the mobile context is limited as they monopolize many resources. The argumentation that supports this method comes from an experiment with a subset of the “KING speech” database. This database provides utterances from speaker conversations over both signal-to-noise radio channels and narrow-band telephone channels. It has been observed that many unlabeled classes of the sample distribution may be encoded as a linear combination of Gaussian basis functions. The authors hypothesized that this model should be computationally inexpensive and easy to implement on real-time platforms. However, the main drawback of their method comes from the initialization of the training process as several parameters such as the mean, covariance and prior of each distribution must fit the data. Such a process may be achieved through several costly methods like a Hidden Markov Model (HMM) or a binary k-means clustering algorithm. In that sense, although the identification process may certainly be efficient when used in a mobile device context, the training phase would probably be computationally overly expensive [7].

A second text independent method for speaker authentication has been described in **.** This method relies on a back-propagation neural network having LPC (Linear Prediction Coefficient) parameters as input features, to predict utterances. The use of the back-propagation method aims to optimize the distribution of weights between neuron layers. Doing so, it becomes possible for the neural network to correctly map arbitrary inputs to outputs. The decision process is achieved by associating each speaker to an utterance. In a database having 25 speech samples of different languages, the identification accuracy that the author was getting was nearly about 85.74%. With this promising achievement, Kumar et al. have concluded that the proposed method would be appropriate and reliable. Nevertheless, we may note that the theoretical complexity of a standard back-propagation neural network training phase is O

(nmh^k*oi) where, n are training samples; m refers to features; k are hidden layers, each containing h neurons; o refers to output neurons; and i is the number of iterations. This suggests that the computation time is still overly expensive considering the limited capacity of mobile devices [8].

Another text independent method for speaker authentication has been proposed by Nair and Salam. This method resorts to both LPCs and LPCCs to compare their strength. The use of the Dynamic Time Warping (DTW) algorithm allows to decide about the best option. The TIMIT (Texas Instruments and Massachusetts Institute of Technology) speech database was used for the experiment. This corpus of American-English Speakers (AESs) counts 630 speech signals. The achieved accuracy is around 92.2% with LPCs. With derivative cepstral coefficients, it climbed to 97.3%. As expected, the association of LPCCs to the DTW algorithm offers an accurate and reliable solution. Since DTW requires a quadratic complexity both in terms of time and memory usage i.e.,

O (n^2))**,** it appears that it may not be the most suitable solution to achieve speaker authentication, directly on the mobile device. Nevertheless, real speaker authentication scenarios usually imply few distinct samples [9].

On the other hand, Brunet et al. have proposed TISA (Truth in Saving Act) method dedicated to mobile devices. This method starts by extracting MFCC features from speech samples. With this information, a Vector Quantization (VQ) method allows to construct a reference model. Euclidean distances between stored centroids and tested samples enable to accept or to reject the attempt based on a given threshold. For their personal database, samples for training and testing were collected with a mobile device. Being implemented as a stand-alone biometric system, the Equal Error Rate (EER) was the only performance indicator [10].

4.Literature Survey

This section consists of the research works that have been done by various authors in the field of Voice Biometrics System.

S. No	Name of Author	Year	Work Done	Problems
1	Kinnunen et al.	2017	Study on old and new SR system	Some trainings are available which are unmatched handsets
2	Saquid et al.	2017	Widespread survey on ASR systems	Low protection scheme in biometric technology
3	Togneri et al.	2018	Outline on text – independent systems	The authors only measured the SI system based on missing data.
4	Tirumala et al.	2018	Comparative study of old SR system versus new recognition system.	Need of alternative for the better representation of audio and video parts
5	Hamidi et al.	2019	Explains the foundation on optimal features and the feature parameters.	Not able to control the feature capacity
6	Irum et al.	2019	Consolidate feature extraction methods	Required performance improvement for the feature extraction methods

7	Vestman et al.	2019	A study on ASR based on Deep learning	Most feature extraction techniques needs be more relevant and exact.
8	Singh et al.	2020	Short utterance effect investigation	The approaches that are used for ASR systems is not accurate.
9	Manjutha et al.	2020	The ASR systems' effectiveness was found out using the various features	The authors found out very hard to find out the coefficient weight of the system
10	Sujaya et al.	2021	Feature extraction methods were used to take out the required feature only from the input	The already proposed Bottleneck feature extraction was worse as compared to the feature extraction methods that we are using
11	Desai et al.	2022	Deep speaker feature training	Suggestions on linguistic and acoustic systems were given which then will be improved.
12	Lawson et al.	2022	CNN based ASR systems was made and introduced.	The system was getting error while doing the experiment

5.PROBLEM STATEMENT

The ASR systems developed earlier faced several challenges viz inter-mixing of voices thereby creating noise, low accuracy of result (voice identification) and replay attack. Due to these drawbacks, the ASR system has been losing popularity in usage and authentication. Also, with these drawbacks, the earlier ASR systems were more vulnerable to duplication and spurious attacks resulting as a loss and danger to the society and mankind.

To overcome these challenges, we have introduced the following – (i) filtration of the voice; after removal of text; to extract the accurate voice features by removal of the unwanted noise mixed with it. (ii) We have introduced double authentication mechanism by the use of a pass-phrase after the filtered voice sample is matched initially. (iii) To make ASR system robust against the replay attack, we shall use a randomly prompted text/phrase for authentication and for this we will design and implement “text-independent biometric authentication system.” In addition to this, another objective of this work is to describe various speech segmentation and clustering techniques that can used in speaker recognition for more accurate result.

6. PROPOSED METHODOLOGY

In this, we are proposing a new verification system for the mobile devices based on the speakers which are text independent. This system is fully automatic as all the processes is operated on mobile devices. So, our method does not require any kind of server/platform for the conduction of the processes. The architecture of our system is mentioned in figure 1 . The first step contains the extraction of a selected set of individual voice features from an audio signal coming from the speaker, in order to build a dataset. In second step, it takes those datasets as input and performs a training session. For the training session, it uses "Naive Bayes Classifier." The third process calculates the validation decision and gives results in the form of true or false.

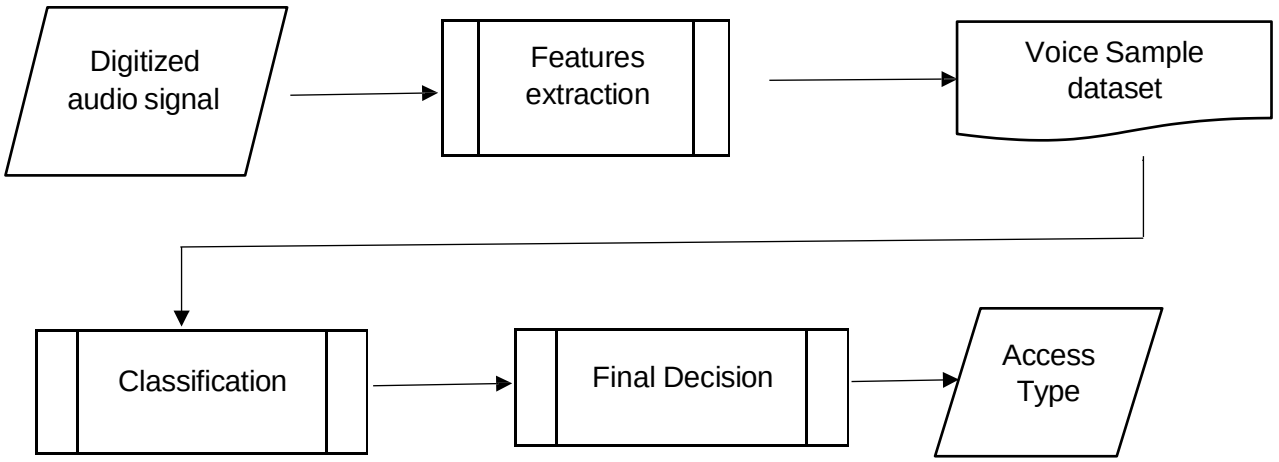

Fig 1: Flowchart of the proposed methodology

The basic processes that are involved for the proposed model is given below: -

A.) Input: Every audio file will be recorded with an integer PCM encoding format which is of 16 bits that will use two channels. The sampling rate of the audio file that will be used here is approximately 44.1 kHz.

B.) Pre-Processing: This section will describe the pre-processing phase of our approach. This phase contains two steps viz voice activity detection and audio normalization.

(i) Voice Activity Detection- The first step of the pre-processing phase consists of the voice of the user that the user will give and the voice will be divided into segments for keeping it safe. For getting this, we will define a fixed threshold value that will be used to get. We use that threshold to identify the sections of the input signal that we need to remove.

(ii) Audio Normalization- After the removal of the unwanted voice from the input, audio normalization process is carried out [11]. The goal of this process is to see a consistent change in the voice that will be in the form of graph. The original information is not impacted as the complete signal is changed which results in proper process of the voice.

(iii) Feature Extraction - The task of separating distinguishing qualities from an audio stream comprising a speaker's voice is not simple. The voice is observed as a particularly specific

signal that contains extensive information about the speaker. As a result, one of the main components of speaker identification and authentication systems is the extraction of information from the speech. In our strategy, we suggest utilising Linear Prediction Cepstral Coefficients (LPCCs) to extract features [12] . Such coefficients are obtained directly from the linear prediction analysis, which seeks to determine the pertinent properties from a speech signal. While maintaining a good computation speed, we can provide extremely precise estimates of the speech parameters by using LPCCs .

(iv) Classification - Several classification techniques have been applied for classification of SR i.e., GMM, ANN, etc. Naive Bayes classifiers are well known for being quick, highly effective, and simple to use. The classification algorithm used in the method is supervised and statistical, and it calculates the conditional probabilities of the various classes given the value of the characteristics. The class with the highest conditional probability is finally chosen.

(v) Decision Making - Decision-making process (granting access or not) is very important stage in our system. It can generate two types of errors. At first, it may result in a false negative error, which means that the system fails to identify a genuine user. Or, it may result in a false positive error, which means granting access to a non-authorized user [13]. While a false negative authentication does not compromise the security or the privacy of the user's data, it constitutes a huge source of prevention. In that case (false negative), the authentication process must be re-done, or a fall-back mechanism (i.e., a PIN number) must be used. The process is shown in Fig.2 below.

Based on the processes involved in identification and verification, it is obvious that verification is quicker than identification. Most of the time, identification is done first to discover the best match, then only verification is done to get to a certain conclusion, according to the research done in this area. Consider the following scenario: "If a voice sample of a suspected assaulter is captured, this voice sample is matched with the previously formed voice database and tried to find the best match of this sample (that is identification) after that verification is performed, then gives a conclusive result declaring true or false and the best match voice is belong to that assaulter or not." This can be justified by taking this example.

We can say that all the five mentioned processes will allow a certain flexibility in the access of the mobile devices as well as the ASR systems. Indeed, we provide a means to define an appropriate level of restriction since each user has many different considerations according to what piece of information, he/she has on his/her phone that is important as regards confidentiality or not. Users are thus able to tune the system to make the decision-making process restrictive and to best fit their personal needs.

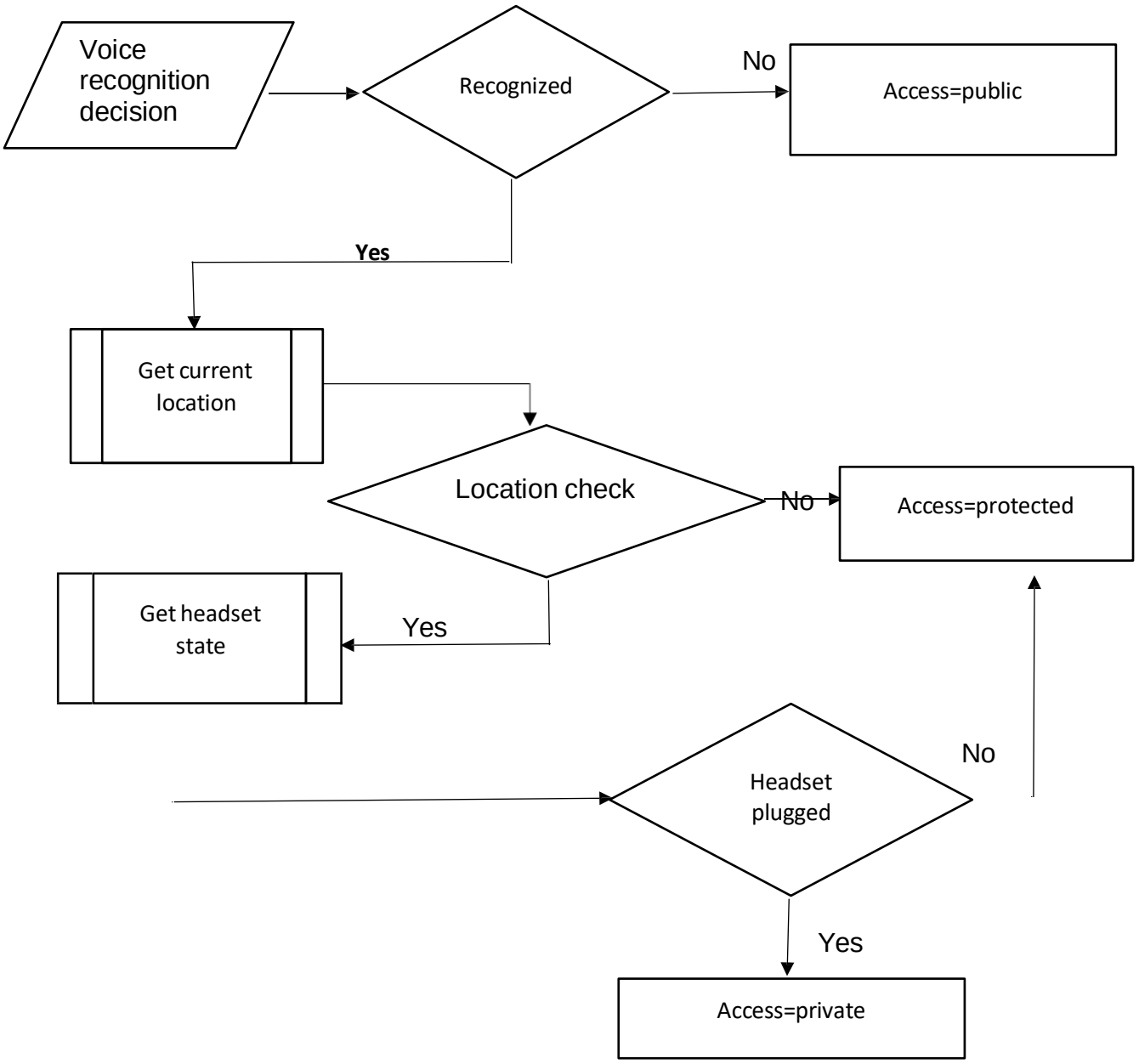

Fig 2: Flowchart of the decision- making process

7. CLASSIFICATION OF SPEAKER RECOGNITION(SR)

ASR can be divided into a variety of classifications based on the recognition criteria. The following subsections provide a detailed description of the various speaker recognition techniques:

(A) Identification, Verification and Diarization

As compared to other categorization criteria, this level of classification is the most important. The most basic and effective methods for preventing unauthorised access to computer systems are widely regarded as Speaker Identification (SI), Speaker Verification (SV), and Speaker Diarization (SD). Each of these speaker recognition subdomains is explained as follows:

(i) Speaker Identification (SI): Using a list of recognised speaker voices, speaker identification selects the precise speaker. It is a way to find someone using the different utterances that are classified in the database. This method compares a specific utterance in a 1: N match against N templates.

(ii) Speaker Verification (SV): To verify a particular identification that the speaker has claimed, SV works with the voice. With SV systems, on the other hand, the acquired traits are solely connected to the speaker's stated stored traits. A speaker's speech is compared to a single template in a 1:1 match.

(iii) Speaker Diarization (SD): To identify everyone in a voice with several speakers, SD divides the voice into homogeneous chunks. Speaker recognition systems must include it. It is useful for comprehending conversational content, video captioning, and many other crucial areas.

The following section explains whether “Text -dependent” or “Text Independent” should be the recognition criteria for SI, SV, and SD systems.

(B) Text Dependent and Text Independent Recognition

For SR, text dependence is a different classification level. The text that the speaker spoke while classifying it was used to make this classification. The two text-based speaker recognition subdomains are described as follows: -

(i) Text – Dependent: It is a speaker recognition job, such as verification or identification, is referred to as speaker recognition if only a portion of the vocabulary used during testing is the same as that utilised during enrolment. The limited vocabulary makes the procedure more challenging in a scientific and technical sense. Yet, the procedure also enables speedy enrolment (or registration) and testing sessions to deliver an accurate response.

(ii) Text – Independent: Verifying the speaker's identity through text independent speaker verification entails doing it without imposing any restrictions on the speech's content. Since the user has more freedom to speak to the machine, it is more useful than SV. However, longer period of training and testing are required to achieve good performance and accuracy.

More than 28 literature surveys on SR have been written by the authors in a thorough survey, with nine of them being specifically addressed. The influence, difficulties, and present trends

of speaker recognition systems are not specifically included in these studies . On the entire field of speaker recognition, there has not been a thorough study done in the recent past. In 2018, a renowned author conducted a thorough literature review on major feature extraction techniques in her previous studies. The report made many recommendations based on the analysis and requirements, to identify and implement the significant feature extraction techniques that have been already used in the last six years .

The best feature for ASR was studied as a foundation using the feature extraction process, and the authors offered solutions for the speaker identification problem. Deriving feature extraction approaches and architectures led to the discovery of the most well-known and productive feature extraction techniques over the preceding six years. Finally, they discussed a few difficulties in the speaker recognition field. Another author named Diksen conducted a thorough analysis of "speaker-specific information extraction techniques." According to how well the theories stood up to channel mismatch, additive noise, and other flaws like vocal effort, emotion mismatch, etc., they were divided into three classes . The authors' main areas of interest were speaker-specific information in degraded circumstances and short-time feature extraction. In their later works, authors Togneri and Pullella explained how to handle missingdata in order to improve the robustness of the SR system and provided the real speaker identification example. In their work, they discussed the modelling approaches, feature extraction strategies, and model classification of numerous speakers.

ASR (Automatic Speaker Recognition) is a process that uses a machine to identify a speaker from his or her spoken words or sentences. In several areas, including crime investigation, access control, voice-based banking, and authenticity, this technology helps to maintain security. One of the key tasks in the development of an ASR system is the extraction of speech signal characteristics. Feature extraction techniques' main task is to separate voice features from the spoken signal. Every individual has their own distinct voice qualities that can be utilised to identify them.

With a focus on deep learning-based techniques like SI, VI, and SD, Bai and Zhang covered several significant speaker recognition sub-domains. They provide a thorough overview of modern, recently released deep learning-based feature extraction methods and ASR algorithms in their paper. The speaker recognition industry has also intermittently introduced a few other polls. These surveys do not adequately support the speaker recognition domain, hence further research is necessary.

8. Performance Evaluation of Voice Biometric System

The following variables affect the performance of a Speaker Recognition system.

(a) FAR (False Acceptance Rate): - Percentage of identifying cases where an illegal person is mistakenly accepted.

(b) FRR (False Rejection Rate): - It measures the difference between the total feature and a feature that was incorrectly rejected in relation to the total feature.

(c) ROC (Relative Operating Characteristic): - The distinction between the FAR and FRR is represented graphically. The acceptance or rejection of the matching algorithm's output determines how well the speaker recognition system performs. A threshold that

determines the recognition value close to it serves as the foundation for the matching algorithm's results.

(d) EER (Equal Error Rate): - The equal error rate (EER), which is used in ASR systems, is a technique for assessing system performance. To determine the EER, often multiple testing samples are required. The FRR and FAR are equivalent at EER (also known as THRESHOLD). EER swiftly equalises the accuracy of any system with the aid of various ROC curves. More accurate is the system with the lowest EER. EER is calculated as per the following formulae **EER = FAR + (FRR * FRR).**

The graphical representation below gives a better understanding of FAR, FRR and EER.

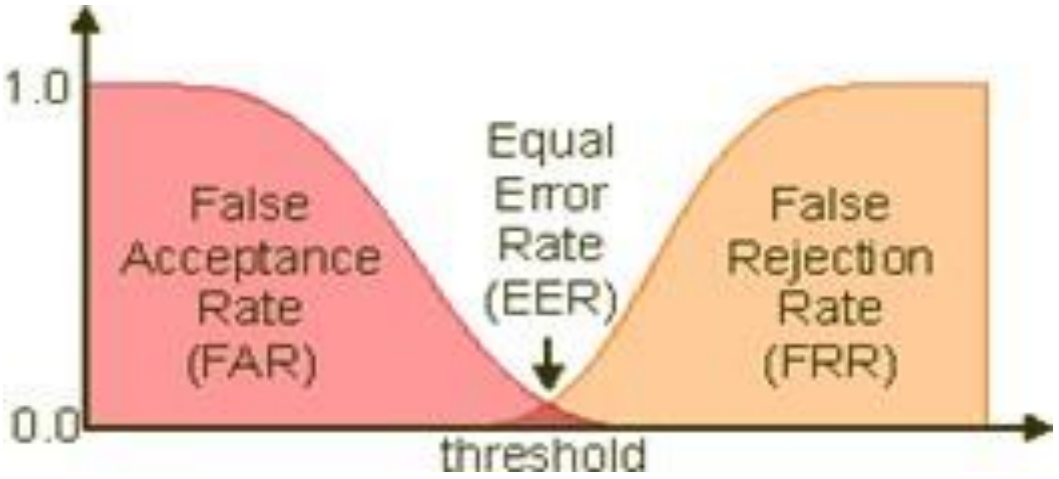

Fig 3: FAR, FRR & EER [14]

9. PROBLEMS WITH ASR SYSTEMS

Methods for voice detection face several difficulties. Data-driven solutions frequently run into the issue of intra-speaker variability. In both text-dependent and text-independent speech recognition tasks, this issue frequently occurs. The basics of speaker identification are discussed in the first part. Then an explanation of the technological constraints follows.

(A) General Challenges: Tasks required for Speaker Recognition system are challenging in both text-dependent and text-independent models.

(i) Dependence on data: - Despite being practical in nature, a large portion of Speaker Recognition's techniques are data-driven. It takes a huge amount of expertise to train the background algorithms. The database has already required a lot of human labour to structure and organize [15].

(ii) Intra-speaker variability: - There can be a significant amount of variation as the same speaker does not always deliver the very same speech in the same manner every time.

We have spoken about two different difficulties that make speaker recognition tasks very difficult and gives inaccurate result [16]. This variability in result is based on two aspects, viz-

(a) Variability based on Conversation: - It represents various scenarios, including verbal exchanges with other people, systems, or issues that involve a specific language or accent. It comprises dialects spoken between people, monologues.

(b) Variability based on Technology: - There are additional challenges with the location, timing, ambient, electromechanical, and data quality, including duration, sampling rate, recording quality, and audio compression.

(B) Technology Challenges: The underlying algorithms and the technical difficulties of speakerrecognition architectures are closely related. Few technological issues are listed below: -

(i) Limited data and constrained lexicon: - Modern industrial bootcamps frequently consist ofrepetitive repetitions of the enrolment lexicon. A single repetition of a tiny section of the recorded vocabulary for 4-5 seconds of total speech input is used to determine the trial's duration.

(ii) Use of Channels: - The fact that the most successful app users use a range of phones, including landlines, payphones, cordless phones, cell phones, etc., is not surprising. This worsens their effect on the effectiveness of channel utilisation. A "cross-channel endeavour" isa measuring interval that starts during the speech verification period on a different channel thanit did during the training period. In this domain, it is essential to improve SR systems.

10. RESULT

Each test was repeated 10 times for eachspeaker, i.e., for a test of 5-word length, 10 different speech files of 5-word length from the same speaker were tested. Total number of successes or failures were calculated by adding test results from all the speakers. Each test was repeated for 5-word, 10-word, and 15-word test length.

10.1 Authentication Performance Evaluation

In case of verification application, A UBM was trained by merging short speeches of many speakers for 6.13 minutes duration. The UBM acted as an alternate (*impostor*) model in authentication application. Decision to accept or reject the claim wasbased on likelihood ratio of the match score with claimed identity and alternate model (UBM).

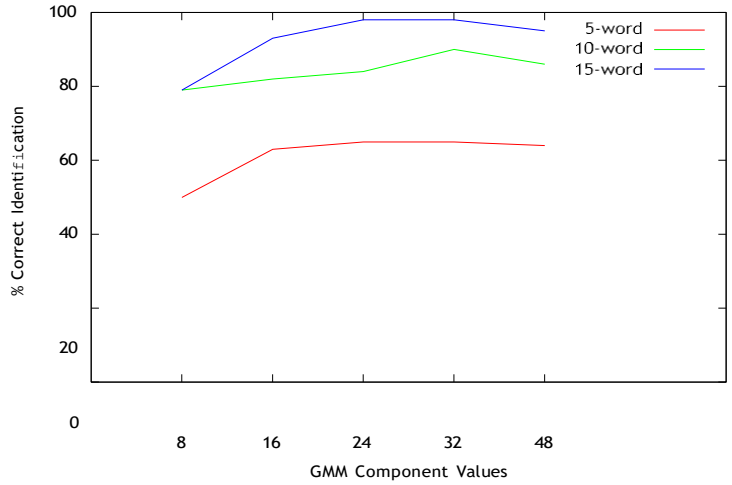

Figure 4: Effect of *GMM* Order and Test Length on Identification Performance.

Test length	Training Duration				
	1 minute	2 minute	3 minute	4 minute	5 minute
5-word	72	63	73	68	65
10-word	87	84	87	86	90
15-word	90	91	95	97	98

Table 11.3: Effect of Training Duration and Test Length On Identification Performance.

Experiments for evaluation of the authentication performance were conducted on 9 registered male speaker models. For calculating FRR,claimant was an authentic user. Ideally, the claim of an authentic user should not be rejected. FRR is calculated as:

FRR = (Number of test samples incorrectly rejected/Total number of test samples) * 100

FAR= (Number of test samples incorrectly accepted/Total number of test samples) *100

Total Error Rate (TER) is calculated as:

TER= (%FAR + %FRR) / 2

Test Length	% FAR	% FRR	% TER	% Auth. Accuracy
5-word	6.11	18.57	12.34	87.66
10-word	4.44	12.50	8.47	91.53
15-word	5.72	4.44	5.08	94.92

Table 5: Authentication Performance: FAR, FRR, TER, and authentication Accuracy.

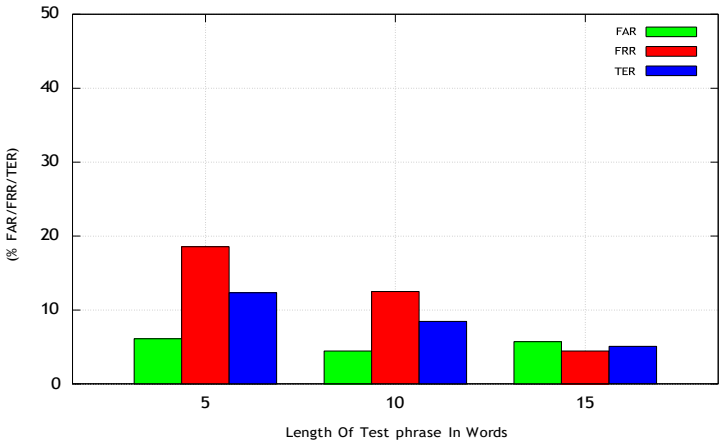

Figure 6: Authentication performance: FAR, FRR, and TER.

11. CONCLUSION

Speaker recognition (SR) is a highly praised study area that has been extensively studied and incorporated into many systems to identify or validate speaker. However, only a little amount of research work has been done, and most of it is now out of date. The foundations of the domain, feature extraction methods, datasets, architectural designs, performance evaluations, and challenges encountered are just a few of the topics covered in this paper. The main goal of the paper is to outline the fundamental concepts and ongoing research in the ASR field.

The report also compiles the most recent ASR system implementations, studies them, and offers a comparison based on the afore-mentioned factors. The report strongly supports the defaults and variants of ASR systems that are now in use and that new researchers can use to quickly adapt to the notions of the research domain.

Some significant forensic speaker recognition topics has been covered in this report. A forensic technique's guiding principle is the idea of uniqueness. Forensic linguistics (also known as voice biometrics) is a field of research and engineering that studies the legal issues associated with digital evidence.

ASR systems need to be more accurate because they are used in so many crucial and secure situations, such as speech indexing, sensitive medical data, online transactions, fraud, and access control. ASR system is expected to have a very wide range of applications in near future.

12. FUTURE SCOPE

Our system is not optimized or examined by considering effects of speech recognition module on accuracy and speed. In this work, role of the speech recognition module is limited to providing security against replay attacks. As a future work, system's efficiency can be examined and optimized by considering the performance of speech recognition module also. System performance can also be tested and optimized for remote authentication via telephone channel. This work was tested for a small database of 14 speakers. The system should be examined and optimized for a larger speaker data set as well.

References

[1] Voice Biometric Systems for User Identification and Authentication – A Review - Amjad Hassan Khan M. K., & P. S. Aithal

[2] An Overview of Text-Independent Speaker Recognition: from Features to Supervectors - Tomi Kinnunen, Haizhou Lib

[3] A Study on Automatic Speech Recognition - Saliha Benkerzaz, Youssef Elmir, Abdeslam Dennai

[4] An Overview of Text-Independent Speaker Recognition: from Features to Supervectors - Tomi Kinnunen, Haizhou Lib

[5] An Overview of Text-Independent Speaker Recognition: from Features to Supervectors - Tomi Kinnunen, Haizhou Lib

[6] An Overview of Text-Independent Speaker Recognition: from Features to Supervectors - Tomi Kinnunen, Haizhou Lib

[7] A Text-Independent Speaker Authentication System for Mobile Devices Florentin Thullier , Bruno Bouchard and Bob-Antoine J. Menelas

[8] A Text-Independent Speaker Authentication System for Mobile Devices Florentin Thullier , Bruno Bouchard and Bob-Antoine J. Menelas

[9] A Text-Independent Speaker Authentication System for Mobile Devices Florentin Thullier , Bruno Bouchard and Bob-Antoine J. Menelas

[10] A Text-Independent Speaker Authentication System for Mobile Devices Florentin Thullier , Bruno Bouchard and Bob-Antoine J. Menelas

[11] S. Furui, “Fifty years of progress in speech and speaker recognition”, Proc. 148th ASA Meeting, 2004. doi: 10.1121/1.4784967

[12] S. K. Singh, Prof P. C. Pandey, “Features and Techniques for Speaker Recognition”, M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03.

[13] Davis S., Mermelstein P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366, 1980. doi: 10.1109/TASSP.1980.1163420

[14] “Vector Quantizer Encoder: Blocks(Signal Processing Blockset)”, The Mathworks incorporation, 1982-2008

[15] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies", General Aspects of Digital Transmission Systems; Terminal Equipments, International Telecommunication Union (ITU), 1993.

[16] Beigi, Homayoon (2011). “ Fundamentals of Speaker Recognition.”.[Online].Available: http://www.wikipedia.org/wiki/speaker_recognition.