Computers and Electrical Engineering 90 (2021) 107005

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

A review on speaker recognition: Technology and challenges

Rafizah Mohd Hanifa [a,b,*], Khalid Isa [a], Shamsul Mohamad [a]

[a] Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[b] Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia

ARTICLE INFO

Keywords:
Biometric
Open system
Speaker recognition
Text-independent
Feature extraction
Classifier
Machine learning
Adversarial attack

ABSTRACT

Voice is a behavioral biometric that conveys information related to a person's traits, such as the speaker's ethnicity, age, gender, and feeling. Speaker recognition deals with recognizing the identity of people based on their voice. Although researchers have been working on speaker recognition in the last eight decades, advancements in technology, such as the Internet of Things (IoT), smart devices, voice assistants, smart homes, and humanoids, have made its usage nowadays trendy. This paper provides a comprehensive review of the literature on speaker recognition. It discusses the advances made in the last decade, including the challenges in this area of research. This paper also highlights the system and structure of speaker recognition as well as its feature extraction and classifiers. The use of speaker recognition in applications is also presented. As recent studies showed the possibility of fooling machine learning into giving an incorrect prediction; thus, the adversarial attack is also discussed. The aim is to enhance researchers' understanding in the area of speaker recognition.

1. Introduction

The growing interest in security has seen a rise in the use of biometrics. Besides the face, other unique features such as retina, voice, and iris can also distinguish between people. As shown in Fig. 1, biometrics can be classified into two categories: physiological and behavioral [1,2]. The former includes the face, fingerprint, and iris, while the latter includes voice, keystroke, and signature.

Table 1 provides the typical characteristics of biometric technologies and their performances in terms of accuracy, ease of use, user acceptance, ease of implementation, and cost [3,4]. Based on the information in the table, it can be deduced that voice is one of the most useful technology, as it is easy to use, easily implemented, and widely accepted by users due to its low cost. Furthermore, the study by [5] asserted that besides iris, fingerprint, and face, voice is another useful biometric because it provides a comparable and much higher level of security. Meanwhile, the study in [1] stated that voice could be used to differentiate people because each person's voice has some unique characteristics.

In general, any sound produced by humans to communicate meanings, ideas, opinions, etc., is called a voice. In a specific term, voice is defined as any sound produced by vocal fold vibration, which occurs when air is under pressure from the lungs [6]. Voice is the most natural communication tool used by humans. It conveys the speaker's traits, such as ethnicity, age, gender, and feeling.

Research in speaker recognition has been conducted worldwide in the last eighty years but has increased significantly due to the advancements in signal processing, algorithm, architecture, and hardware [7]. Automatic Speaker Recognition (ASR) is a digital signal processing field related to recognizing people's voices. Every individual's voice is unique due to the differences in the shapes of the

* Corresponding author.
E-mail addresses: rafizah@uthm.edu.my (R. Mohd Hanifa), halid@uthm.edu.my (K. Isa), shamsulm@uthm.edu.my (S. Mohamad).

https://doi.org/10.1016/j.compeleceng.2021.107005
Received 27 April 2020; Received in revised form 15 January 2021; Accepted 19 January 2021
Available online 25 January 2021
0045-7906/

vocal tract, larynx sizes, and other parts of human voice production organs [8,9]. The features of voice are dependent on its pace or speed, volume, pitch level, and quality, while the articulation rate and speech pauses rely on the speaker's speaking style [8]. Although there have been several review papers published in the field of ASR, each covers different perspectives. Thus, this paper aims to provide a thorough review of the ASR system and its latest issues and challenges.

This paper is organized as follows: Section 2 presents a detailed explanation of the difference between the terms "speaker recognition" and "speech recognition", as well as highlights the structure of speaker recognition including its feature extraction, classifiers, and models. The taxonomy of speech processing technology involving the types and issues of speaker recognition including the adversarial attack is addressed in Section 3. Section 4 presents the milestones in speaker recognition in the last decade based on the feature extraction being used. The last section concludes the paper.

2. Speaker recognition

In speech processing, speaker recognition and speech recognition are the two applications commonly used by researchers to analyze uttered speech [10]. Before delving further into the structure of speaker recognition, it is vital to understand the difference between speaker recognition and speech recognition. Speech recognition is concerned with the words being spoken, while the speaker or voice recognition aims to recognize the speaker rather than the words [11].

2.1. Speaker recognition vs. speech recognition

Speech recognition is useful for people with various disabilities, such as those with physical disabilities who find typing the words difficult, painful, or impossible, and for those who have difficulties recognizing and spelling words, such as those with dyslexia [12]. Since speech recognition deals with converting audio into text, its effectiveness depends heavily on the language and the text corpus [5]. On the other hand, speaker recognition is to identify the person who is speaking. Pitch, speaking style, and accent are some of the features that contribute to the differences [11]. Speaker recognition technology has been used in various applications, such as biometric, security, and even human-computer interaction. Table 2 summarizes the differences between speaker recognition and speech recognition in terms of recognition, purpose, focus, and application [13].

The advancement in various fields has increased the importance of speaker recognition systems, especially in identifying a person's identity.

2.2. Structure of speaker recognition

Speaker recognition involves the process of finding the identity of an unknown speaker and comparing his/her voice with those available on the database. It is a one-to-many comparison [14]. The basic framework and components of speaker identification, as shown in Fig. 2, consist of two phases: enrolment, also known as training, and recognition, also known as testing [15,16]. The following subsections present the main phases involved in the framework.

2.2.1. Pre-processing

Pre-processing is the first step in speech signal processing, and it involves converting an analogue signal into a digital signal [9,17]. Interference due to noise often occurs during speech recording, causing the performance to degrade. Thus, pre-processing is a crucial and critical step, as improper pre-processing conducted on the recorded speech input will decrease the classification performance [17]. The main objective in the pre-processing stage is to modify the speech signal to be suitable for feature extraction analysis [16,18]. Different methods can be adopted for noise-reduction algorithms, and the two most frequently used are spectral subtraction and adaptive noise cancelation [18]. However, [19] highlighted that the function to be used during the pre-processing stage is very much dependent on the approach employed at the feature extraction stage. Some of the commonly used functions include noise removal, endpoint detection, pre-emphasis, framing, and normalization [16,17,19].

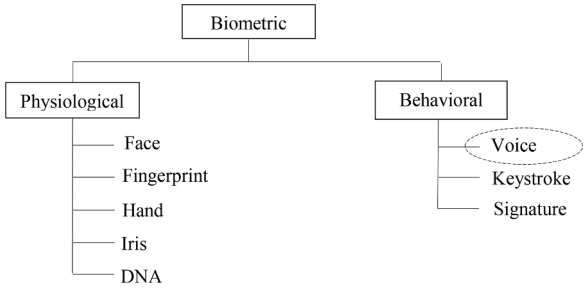

Fig. 1. Types of biometrics: Physiological and Behavioral.

2.2.2. Feature extraction

Feature extraction is a significant issue in the area of text-independent speaker recognition [5]. The basic principle of feature extraction is to extract a sequence of features for each short-time frame of the input signal, with the assumption that such a small segment of speech is sufficiently stationary to allow for better modeling [20]. In other words, the process retains useful and relevant information about the speech signal by rejecting redundant and irrelevant information [8,18]. This phase is vital for the next step, as it affects the behavior of the modeling process. The speaker signal is a dependent speech system whereby the speech signal is analyzed to get less variability and to identify more discriminative features by converting a speech signal to parametric values [21]. Various techniques can be used for extracting speech features in the form of coefficients, such as the Linear Prediction Coding (LPC), the Linear Prediction Cepstral Coefficients (LPCCs) and the Mel-Frequency Cepstral Coefficients (MFCCs) [8,9,17,22]. Table 3 presents a comparative summary of the merits and de-merits of the various feature extraction techniques.

2.2.3. Models and classifiers

Speaker models and classifiers are dependent not only on the features used but also on the task to be addressed. Many factors, such as type of speech, ease of training, and computational and storage requirements, need to be considered before choosing the modeling technique [29].

Modeling techniques are categorized into generative models and discriminative models [30]. Generative models present the distribution of individual classes, while discriminative models learn the boundaries between classes. Fig. 3 shows the categorization of modeling techniques.

2.2.3.1. Generative models. Generative models require training data samples from the target speaker and can take the form of a statistical or non-statistical model to describe the target speaker's feature distribution [29]. As shown in Fig. 3, generative models can be further classified into parametric and non-parametric modeling [31]. A model that assumes a structure characterized by certain parameters is known as a parametric model, whereas, in a non-parametric model, the probability density function is made with minimal assumptions [29].

2.2.3.1.1. Parametric models. Parametric models include the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). The benefits of these models are the efficiency of the data used, which can be extended to evaluate the test data through statistical summaries of the data rather than the data itself [32]. The parametric model's disadvantage is that the structure is restrictive and may not be adequate to model the task [29].

2.2.3.1.2. Non-Parametric models. Template matching is a non-parametric model that consists of a template that is a sequence of feature vectors from a fixed phrase [14]. Non-parametric models include Dynamic Time Warping (DTW) and Vector Quantization (VQ). The advantage of these models is that they are free from assumptions about data generation [8]. Besides, this method does not require any model training [29]. The disadvantage of this model is that the pre-recorded templates are fixed. Thus, variations in speech can only be modeled using many templates, which becomes impractical [33].

2.2.3.2. Discriminative models. These models require training data for both target and non-target speakers to obtain the optimal separation between the different speakers [29]. Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are models in this category, and they are also known as soft computing modeling. The advantage of these models is their flexible architecture and discriminating training power while their disadvantage relates to having to use trial and error in obtaining an optimal structure [29].

2.2.4. Classifier choice

The choice of a classifier is very much dependent on the application and the constraints that influence the classifier choice [34], as shown in Table 4.

2.3. Speaker recognition and its application

Speaker recognition is employed in broad application areas, such as authentication, personalization, surveillance, and forensic due to high public acceptance and accuracy rate, low-cost smart devices, and more effortless installation of software. In the following subsections, some examples of the applications of speaker recognition are presented.

Table 1
Comparison of different biometric characteristics.

Biometric Type	Characteristics of biometric technologies				
	Accuracy	Ease of use	User acceptance	Ease of implementation	Cost
Voice	Medium	High	High	High	Low
Face	Low	Low	High	Medium	Low
Iris	Medium	Medium	Medium	Medium	High
Fingerprint	High	Medium	Low	High	Medium
Retina	High	Low	Low	Low	Medium
Hand geometry	Medium	High	Medium	Medium	High
Signature	Medium	Medium	High	Low	Medium

Table 2
Speaker recognition vs. Speech recognition.

Features	Speaker Recognition	Speech Recognition
Recognition	Recognizes who is speaking by measuring voice pattern, speaking style, and other verbal traits.	Recognizes what is being said and converts them into text.
Purpose	To identify the speaker.	To identify and digitally record what the speaker is saying.
Focus	Biometric aspects of the speaker, such as pitch, intensity, etc., to recognize him/her.	Vocabulary of what is being said by the speaker and turns the words into digital texts.
Application	Voice biometrics.	Speech to text.

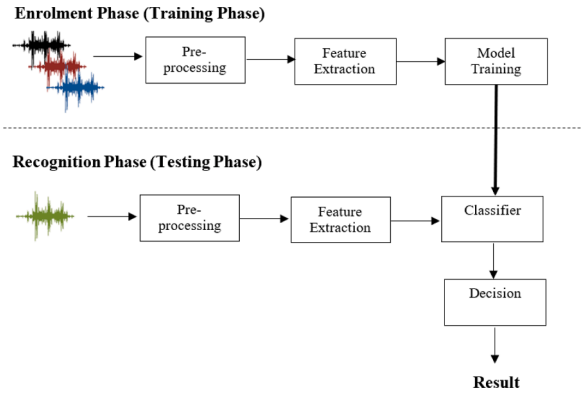

Fig. 2. Framework of a speaker identification system.

Table 3
Comparison of different feature extraction techniques.

Technique	Merits	De-merits
LPC	• Based on basic principles of sound production [20]. • Simple to implement and mathematically precise [19].	• Performance degradation in the presence of noise [20, 23]. • It does not represent vocal tract characteristics from the glottal dynamic; thus, consumes time and computational cost [9]. • Inconsistency with human hearing [24].
LPCC	• Smoother spectral envelope and stable representation as compared with LPC [20]. • Feature components are decorrelated [19, 25].	• Linearly spaced frequency bands, which is inadequate [19,25]. • Sensitive to the quantization noise [23]. • The performance is degraded in case of using insufficient order [23].
MFCC	• More information on lower frequencies than higher frequencies due to mel-spaced filter banks. Hence, behaves like human ear compared with other techniques [20, 26]. • Captures the main characteristics of phones in speeches with low complexity [27].	• Based on Short-Time Fourier Transform (STFT) which has fixed time-frequency resolution [26]. • Low robustness to noise [19, 28].

2.3.1. Authentication

Authentication is one of the most popular biometric applications as it allows the users to identify an individual based on his/her voice. Usually, to authenticate the speaker, a combination of techniques is used, such as a password or facial recognition [22]. This biometric authentication could reduce the problems of misused identity and is also more convenient than using a Personal Identification Number (PIN) or password, which can be easily forgotten. In 2015, Gomar invented and patented a speaker recognition system for authenticating a mobile device user [29]. The system produced a biometric voiceprint from the user's speech utterances which are stored in the mobile device when it met a quality threshold. For verification purposes, the user will need to utter an attribute from the biometric voiceprint in order to be an authorized user.

2.3.2. Personalization

A personal digital assistant used to be a luxury reserved for the wealthy, but it is no longer the case nowadays as everyone can have a digital assistant. Siri, Alexa, Cortana, and Google Now, to name just a few, are examples of software that can help us comprehend and carry out our capricious spoken commands. We can do anything with them, such as planning for meetings, event scheduling, shopping,

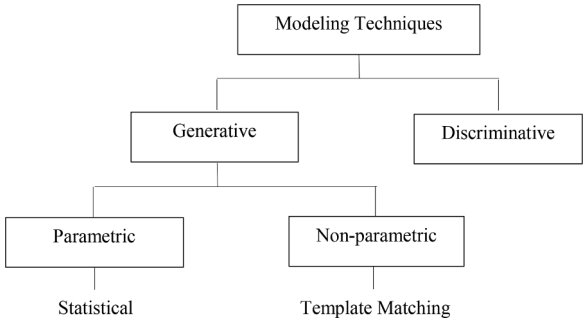

Fig. 3. Classification of Modeling Techniques.

Table 4
Application constraints that influence the classifier choice.

Constraint	Explanation
Level of user cooperation	• If the user is cooperative, the system can ask the user for additional input speeches to boost the performance. • If the user is uncooperative, the system must take more of a passive role, which means the system has no control over the data.
High recognition/ detection accuracy	• This requirement is important in banking account access, as only an authorized user can access his/ her account. • A text-dependent system is applicable, as it offers a higher performance compared with the text-independent technique.
Expected channels	• Channel conditions include the type of microphone used to record the speech, the way the speech is encoded/transmitted and whether the speech contains noise or is free from noise. • If the application must deal with a variety of channel conditions, the classifier could employ channel compensation to boost performance.
Amount of speech data available for enrolment and detection	• If more data are available, then the classifier has higher level of information, which helps in getting better classification.
Available computational and memory resources	• Embedded devices have limited amounts of processing power and available memory. • A cell phone has very limited capabilities, which uniquely constrains the speaker recognizer.
The output of the system	• The output of the system is dependent on the end-user. • As for forensic applications, the system must return word usage and phonotactic information. • Furthermore, the type of output may need to be a hard decision, a human interpretable score, or a relative score.

Table 5
Types of speaker recognition.

Type of Speaker Recognition	Description
Speaker Identification	The task of speaker identification is to classify an unknown voice spoken anonymously as belonging to one of a set of N reference speakers [40].
Speaker Verification	The task of speaker verification is to decide whether the unknown voice belongs to a specific reference speaker with two possible outcomes: to accept the reference speaker or to reject the impostor [40].
Speaker Detection	The speaker detection's task is to mark the target speaker's speeches correctly when the target speaker is presented to the system together with the testing speeches [39].
Speaker Segmentation	The task of speaker segmentation is to find the points where the speaker changes when an audio stream is presented to the system [41].
Speaker Clustering	The task of speaker clustering is to cluster correctly many utterances presented to the system and usually the task is done online [41].
Speaker Diarization	The task of speaker diarization is to split the audio automatically into speaker segments and determine which segments are uttered by the same speaker [41].

etc. In other words, multi-tasking can enhance the efficiency of the works, hence saving our time.

2.3.3. Surveillance

Surveillance is mainly important for security agencies to collect important information, such as electronic eavesdropping of telephone and radio conversations [21]. A filter mechanism is needed to find the relevant information, such as recognizing the target speakers who are of interest to the service. Parole monitoring is another application, which calls parolees at a random time to verify that they are in the restricted area [35]. According to [31], it is also used in applications for remote time and attendance logging and

prison telephone usage.

2.3.4. Forensic

Speaker recognition can also be applied in forensic. This can be done if there is a speech sample recorded during the crime, and the suspect's voice can be compared for voice sample matching. The result can prove the identity of the criminal and discharge the innocent during a court case.

3. Speech processing technology

Speech signal processing technology has become a popular communication technology, as many applications use speech to enhance everyday human life. In digital signal processing, ASR is an essential tool for recognizing people based on their voice [2,4]. Human speech can provide much information as the human voice forms a vital characteristic of an individual [7,29]. Accent, language, speech, emotion, gender, and the speaker's identity are some of the information contained in the human voice [2,36], as shown in Fig. 4.

The field of speaker recognition has gained more attention lately. Although researchers have been working on speaker recognition in the last eight decades, advancements in technology, such as the Internet of Things (IoT), smart devices, voice assistants, and smart homes, have made it popular [5]. Fig. 5 presents the detailed taxonomy of speech processing [37,38].

As illustrated in the figure, the domain of speech processing is divided into three major categories: analysis/synthesis, recognition, and coding. Recognition can be further divided into three parts: speech recognition, speaker recognition, and language recognition. Since the focus of this paper is on speaker recognition, the following subsections will discuss this in further detail.

3.1. Types of speaker recognition

Speaker recognition can be further classified into speaker identification, speaker verification, speaker detection, speaker segmentation, speaker clustering, and speaker diarization [39], as shown in Fig. 6. A brief description of these categories is given in Table 5 to help in understanding their differences better.

Research in the field of speaker recognition has recently increased. It can be further classified into a text-dependent system or a text-independent system and an open set system or a closed set system [8,15], as shown in Fig. 7.

In a text-dependent system, the same text is being spoken both during the training phase and the testing phase, while in the text-independent system, there is no constraint on the text being spoken, which makes it more convenient for the speakers [2,4,8]. Thus, the text-dependent system's training process is much faster since it has a fixed set of inputs to validate. But its limitation is the speaker's inconvenience of having to utter the same words each time. In contrast, the text-independent system's training phase is more prolonged, as the model does not consider what is being spoken but instead tries to convert the audio to feature vectors to identify the speaker correctly [5,29].

An open set system refers to a system that does not limit the number of trained speakers, and the test speakers may comprise other than the trained speakers. However, in a closed set system, the unknown voice must come from a set of known speakers [8,15,29].

Speaker recognition is not an easy task as many factors create variances in the speech signals during training and testing sessions, such as changes in people's voices due to time, health conditions, speaking rates, etc. [16].

3.2. Issues in speaker recognition

Although human voice recognition may be perceived as easy, such as in recognizing a person's voice on the phone, the implementation of speaker recognition systems is challenging. There are many speaker recognition issues and among the most challenging issues in implementing reliable speaker recognition systems are variability and insufficient data [7,17]. Besides those two problems, [2] and [14] further identified background noise as one of the issues that can also influence speaker recognition's performance. The problems may be related to either the speaker or technical errors [12]. In the following subsections, a detailed discussion of each case is

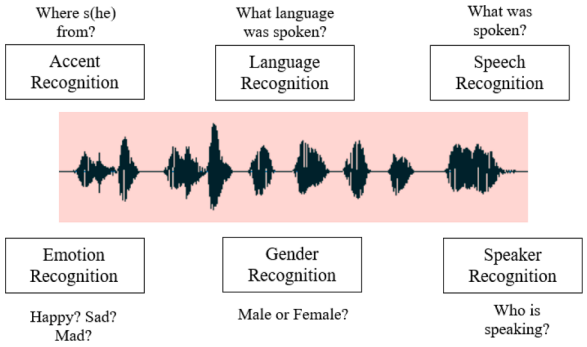

Fig. 4. Some of the information contained in spoken language.

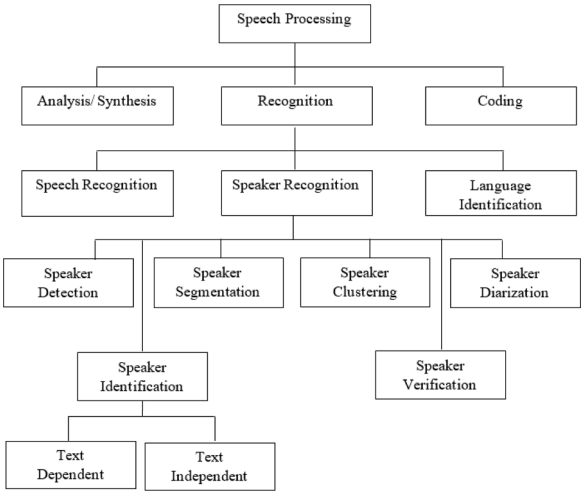

Fig. 5. Speech Processing Taxonomy.

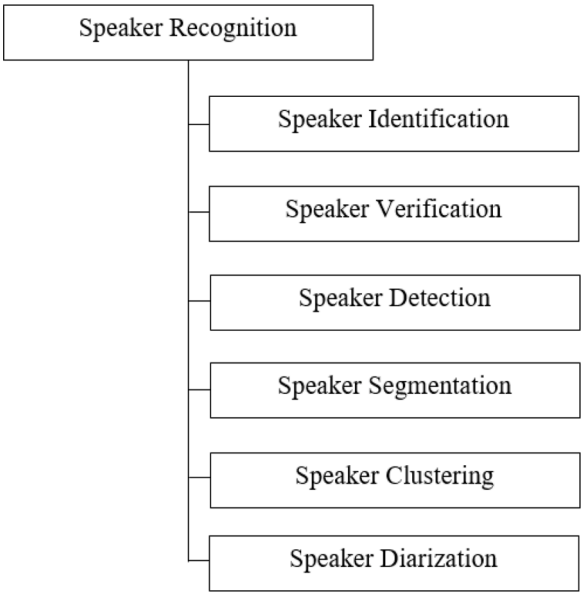

Fig. 6. Categories of Speaker Recognition.

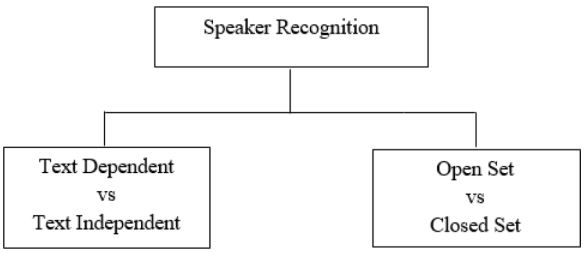

Fig. 7. Parts of Speaker Recognition.

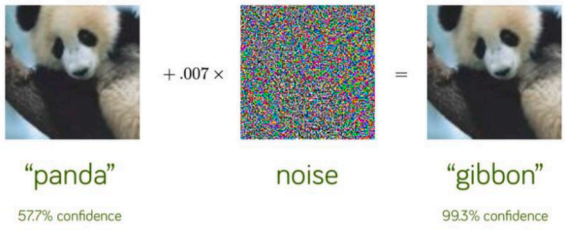

Fig. 8. Adversarial attacks pose. (Source: https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa).

presented.

3.2.1. Issues on variability

The main factor affecting the speaker recognition system's performance is voice variability, also known as session variability, which can be further classified as intra-variation and inter-variation [22]. Intra-variation occurs due to various factors, such as emotions, rate of utterances, mode of speech, disease, the speaker's mood, and the emphasis given to the word at a moment [42]. On the other hand, inter-variation exists due to anatomical differences in the vocal organs and learned differences in speech mechanism [22,42]. According to [7] and [22], the most extreme variations occur during the training and testing sessions, whereby during the training sessions, the speaker utters in a clean environment. In contrast, during the testing session, the speaker speaks in a noisy condition. Hence, [4] suggested that voice recording should be conducted at some interval to avoid changes in the speaker's voice. On the other hand, [2] and [17] further highlighted that speech signals might also change due to different transmission channels, such as the different types of microphones and headphones used during the recording of speech utterances. Thus, this led to the model having a mismatched condition.

3.2.2. Issues on insufficient data

Insufficient data refers to the unavailability of sufficient data to train representative models to reach an accurate decision. It represents a serious and common problem, as most applications require systems that operate with the smallest practical amount of training data recorded in the fewest number of enrolment sessions [7]. The work in [2] highlighted that the problem on this issue is the tendency to assume that the test data distribution matches the distribution represented by the speaker model. The system works if the speaker model is well trained based on long enrolment speech and enough test speeches. In her paper [4], Singh claimed that the system's performance depends on the quantity of training data. According to [14], there is no evidence that the voice sample length provides better results. On the other hand, [2] not only believed that a sufficiently large amount of speech data for model training is necessary but also recognized that, in reality, it is difficult for a system to collect long utterances, or that the user does not speak for long enough. Thus, this led to speaker recognition being a hot topic for further investigation.

3.2.3. Issues on background noise

Background noise is another issue being highlighted by [22] as problematic because during training, the speaker often speaks in a clean environment. In contrast, during testing, the speaker speaks in a noisy condition. The background noise is a significant factor that impacts accuracy in speaker recognition. Accuracy is high for clean samples and low for noisy samples with no impact of babble noise [14]. The recorded speech wave that contains background noises such as white noise, music, etc., has the most significant effect on speaker modeling. It disturbs the evaluation test and degrades the performance of the speaker recognition system.

3.3. Adversarial attacks

Machine learning (ML) and deep learning (DL) have yielded impressive advances in many fields. Unfortunately, recent studies have shown the possibility of fooling machine learning into giving an incorrect prediction.

Fig. 8 shows the adversarial attacks pose where the panda was initially classified correctly. With the addition of noise, the model changed the resulting prediction into another animal, i.e., gibbon, with higher confidence. Although the initial and altered images were the same, they appeared differently to the model. The changes made were the result of the threat of adversarial attacks. If we cannot perceive the difference, we could not tell that an adversarial attack has occurred. Hence, it was difficult to tell if the result was correct or incorrect, even if the model was altered.

Adversarial attacks are categorized into targeted attacks and untargeted attacks [42]. In targeted attacks, the target class will make the target model misclassify the image into something other than the original class. Unlike the targeted attacks, untargeted attacks have no intention of the target class. The aim is to make the target model misclassify by predicting the adversarial image into a class other than the original class.

The research work in [43] constructed targeted audio adversarial examples on automatic speech recognition. They were able to turn any audio waveform into any target transcription with 100% success by adding only a small distortion. They presented proof that audio adversarial examples have different properties from those on images. Another important finding was made by [44], where they integrated the command voice into songs (CommanderSongs). They claimed that their approach was the first to generate attacks

against the DNN automatic speech recognition system. Without any human noticing, the command voice was executed while music was being played over the air. The study in [45] launched a practical and systematic adversarial attack against x-vectors, the DNN-based speaker recognition system. In their work, they added an inconspicuous noise into the original audio. Interestingly, their attack fooled the speaker recognition system into making false predictions and even forced the audio to be recognized as any speaker desired by the adversary.

4. Chronology of advancements in speaker recognition

Research on speaker recognition first started in the 1930s. In March of 1932, the kidnapping and killing of Charles and Anne Lindbergh's baby boy led to a research into speakers' speech signals. During the suspected kidnapper's trial, Charles Lindbergh claimed that the voice of the kidnapper, Bruno Hauptmann, was the same as the voice he heard while waiting in a car nearby where the ransom was paid [46]. Frances McGehee, who was inspired by the case, conducted the first academic research on the reliability of ear witnesses in 1937, which later became a topic of interest in forensics and psychology research [46]. The research in this area continues until today. The reasons for active research in this area in the past few decades are due to various choices of feature selection or extraction, modeling techniques, classification, and decision making, as well as the different databases used [47,48]. In the following paragraphs, the research progress in the past decade is discussed in detail based on the feature extraction techniques.

4.1. Discrete wavelet transform (DWT)

Král in [49] studied the parameterization/classification methods for the Czech language. Two Czech speaker corpora were used in the research. The first Czech corpus contained speeches of 10 native speakers created in laboratory conditions to eliminate undesired effects (e.g. background noise, speaker overlapping, etc.). The second corpus consists of recorded speech in a lesser clean environment, i.e. with some low- level stationary background noise, by 50 Czech native speakers which was created to build a dialog system for Czech Railways. Three DWT with different coefficients (Daubechies with eight coefficients, Symlets with 14 coefficients, and Coiflet with three coefficients) were used and evaluated. The Gaussian Mixture Model (GMM) and the Multi-Layer Perception (MLP) were used as the classifiers for comparison purposes. The results revealed that the best recognition of 99% is achieved with the combination of Linear Prediction Cepstral Coefficients (LPCEPSTRA) with a GMM classifier while, the best configuration of wavelets, which was SYML20 with MLP classifiers, gave an identification rate (IR) of 98% accuracy. The research also showed that using the MLP classifier could reduce training data time to only 30 s compared with GMM, which needed at least one minute. Although the accuracy rate was high, the researcher used a closed set of speakers, which meant the testing voice came from a group of known speakers, which was not practical.

4.2. Mel frequency cepstral coefficients (MFCC)

The research conducted by [50] considered the effectiveness of the fuzzy min-max neural network (FMMNN) as the classifier for closed-set text-independent speaker identification. Since Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used features in speaker recognition, the researchers used them as the features in their work. The uttered words, digits, and sentences in the Marathi language by 50 speakers were used as the database in their experiment. The researchers compared the performances of the two classifiers, FMMNN and GMM. The results revealed an accuracy of up to 99.99% with 15 s of speech utterance when using the FMMNN, while GMM only achieved 97.65% accuracy. As this research also used a closed-set system, it was thus impractical.

The work undertaken by [51] proved that speaker recognition performance could be improved by controlling the noise. Their study was conducted using the TIMIT database, with 100 speakers randomly selected. MFCCs were used as the feature vectors, and the Gaussian Mixture Model-Universal Background Model (GMM-UBM) was used as the classifier. The first set of the experiment evaluated the speaker recognition system's performance under limited data conditions. In the second experiment, the researchers created noisy speech by adding white Gaussian noise to the clean speech. Their work proved that the speaker recognition system's performance in limited data conditions could be enhanced to 80% via artificially increasing the number of feature vectors by adding noise. In contrast, the performance was only 78.20% for limited data and clean speech. They attributed the difference to the relative increase in the number of feature vectors.

The work in [52] used the Continuous Hidden Markov Model (CHMM) to identify Arabic speakers automatically from their voices. Ten Arabic speakers were chosen as the database to evaluate the proposed CHMM-based engine. The CHMM uses the most efficient density function without loss of generality which is the Gaussian density function. Furthermore, CHMM is more accurate than the Discrete Hidden Markov Model (DHMM) as it uses continuous observations to construct the model directly without quantization. MFCCs were used as the feature vectors, and the results showed the IR to be 100% if using text-dependent and 80% if using text-independent. Text-independent is practical as a speaker can be identified from any speech utterance.

On the other hand, [53] proposed a novel hierarchical fuzzy speaker identification method based on fuzzy c-means (FCM) clustering and fuzzy support vector machine (FSVM). FCM was used to reduce the number of training for the audio data and to compute complexity. The FSVM was then used to process the unclassifiable data to make the final decision. For evaluation purposes, the KING speech database with 13-dimensional MFCCs in their first and second derivations was combined into a 39-dimensional vector as input features. Two experiments were conducted for comparison purposes. SVM and FSVM were used as classifiers, and the results indicated that the combination of FCM and FSVM increased the IR from 94.53% to 98.76% compared to using FSVM alone.

The study in [54] presented three approaches of the generalized fuzzy model (GFM) in different roles in their research. The first

model used the Hidden Markov Model (HMM) and the GFM; the second model used the GMM and the GFM, and the last model used the HMM and the GFM with fusion. These three proposed models were tested on the VoxForge speech corpus under clean and noisy conditions and a benchmark database from the National Institute of Standards & Technology 2003 (NIST 2003). For feature extraction, MFCCs were used. The results showed that the model with the HMM and the GFM with fusion gave 93% IR, followed by the other two models with 92% IR.

The work in [55] utilized latent factor analysis (LFA) to deal with channel interference in the speaker's GMM. The algorithm used factor analysis technique to fit the differences between the speaker's characteristics space and the channel space and removed the channel factor in the speaker's GMM. Based on a selection of 38 speakers from the TIMIT speech database and using MFCCs as the feature vectors, the researchers tested the performance with different system modes: Gaussian Mixture Model-Support Vector Machine (GMM-SVM) Linear Kernel, Latent Factor Analysis-Support Vector Machine (LFA-SVM) Linear Kernel, Gaussian Mixture Model-Support Vector Machine (GMM-SVM) Gaussian Kernel, LFA-SVM Gaussian Kernel and GMM-UBM. They found that the LFA-SVM Linear Kernel gave the highest percentage of accuracy (82.84%) compared to the other system modes.

The study in [56] proposed two methods to find MFCC feature vectors with the highest similarity to be applied to a text-independent speaker identification system. Their experiment used 22 speakers selected from the English Language Speech Database for Speaker Recognition (ELSDSR) and utilized the Neural Network (NN) as the classifier. The first method's recognition accuracy was 91.9%, while the second method achieved 93.2%, and the running time was reduced to 42.03% and 20% for the first and second methods, respectively. They claimed that their two approaches could be used for large-scale databases.

The research by [57] explored the text-independent system using MFCCs along with DNN and CNN. The two approaches, DNN and CNN, were adopted for comparison purposes. DNN can provide better noise immunity over GMM, while CNN has been used to better identify patterns in data and scales. Although there was no clear indication of the type of database used in their study, they mentioned recording speakers who used multiple languages (English, Hindi, and Marathi). The voices were recorded using a speaker recognition dataset from openslr developed by Tsinghua University. The results for 50 speakers taken from openslr showed that CNN gave better accuracy, which was 71% as compared to DNN with only 61%. As for the real-world voice samples with eight speakers, CNN again gave a higher accuracy of 75% compared to DNN with only 58%. Thus, CNN as a learning model excelled at identifying patterns in the input and scales much better than DNN.

Neural network-based ASR has shown remarkable power in achieving excellent recognition with enough training data. Unfortunately, the lack of training data prevents ASR systems from performing accurately. Thus, [58] proposed an adversarial few-short learning-based speaker identification framework (AFEASI) to develop robust speaker identification models using a limited training number. Besides employing metric learning-based few-shot learning, they applied adversarial learning to enhance the robustness of speaker identification. Eleven methods were adopted, including seven CNN, one prototypical network (PN), Sincnet (SC), and AFEASI. Accuracy was calculated based on the correctly identified test instance divided by total test instances. Among the methods, AFEASI achieved the highest accuracy of about 0.95 at the setting of 60 s per speaker.

4.3. Temporal Teager energy based subband cepstral coefficients (TTESBCC)

Unlike in [50], the researchers in [59] used whispered speeches with a new feature called temporal Teager energy based subband cepstral coefficients (TTESBCCs) in their study, instead of the FMMNN as the classifier for the closed set text-independent speaker identification. The TTESBCCs were compared with three other feature sets: MFCCs, temporal energy of subband cepstral coefficients (TESBCCs), and weighted instantaneous frequency (WIF). Using a self-generated database of uttered speeches in the Marathi language by 25 speakers and employing GMM as a speaker model, the researchers achieved a higher accuracy rate with the TTESBCCs. The IR for the neutral speech was 98.6%, while the IR for the whispered speech was 55.8%, which were higher compared to MFCCs, TESBCCs, and WIF.

4.4. Normalized dynamic spectral features (NDSFs)

The research work in [60] used a robust spectral feature set called Normalized Dynamic Spectral Features (NDSFs) in a mismatched condition. Their experiment's model formation was based on three different features: Linear Prediction Cepstral Coefficients (LPCCs), MFCCs, and NDSFs. They used two different databases in their investigation. The first database consisted of 100 speakers uttering speeches in English, which was recorded using various sensors. On the other hand, the second dataset used the multi-variability speaker recognition (MVSR) for continuous Hindi speeches generated from the Indian Institute of Technology Guwahati (IITG) database. These two databases allowed them to investigate the mismatch effect. They proved that their proposed features set (NDSFs) were more robust than the cepstral features, such as the MFCCs and LPCCs, with 98% to 100% IR.

4.5. Short-term magnitude spectrograms

The research work in [61] introduced a deep Convolutional Neural Network (CNN)-based neural network speaker-embedding system, called VGGVox, that was trained to map voice spectrograms to a compact Euclidean space, where the distance directly correspond to a measure of speaker similarity. VoxCeleb2 dataset was used for training and validation, while the VoxCeleb1 dataset was used for testing. Two trunk architectures were used in their work: VGG-M and Residual-Network (ResNet). ResNet-50 trained in VoxCeleb2 gave an EER of 3.95% and a cost function of 0.429. For benchmark purposes, VoxCeleb1-E (using the entire set) and VoxCeleb1-H (within the same nationality and gender) were used. The results showed that ResNet-50, which was tested in

VoxCeleb1-E, had an EER of 4.42%. On the other hand, ResNet-50 tested in VoxCeleb1-H, had an EER of 7.33%.

4.6. X-vectors

The research conducted by [62] used the National Institute of Standards and Technology Speaker Recognition Evaluation 2018 (NIST SRE18) in their work, consisting of telephone speeches recorded from the Call My Net 2 (CMN2) corpus and videos extracted from the VAST corpus. They explored Time Delay Networks (TDNN), Extended TDNN (E-TDNN), Factorized TDNN (F-TDNN), and Residual Network with 34 layers (ResNet34). The results showed that E-TDNN x-vectors were the best single system with an EER of 11.1% in VAST and 4.95% in CMN2.

Voice Comparison and Analysis of the Likelihood of Speech Evidence (VOCALIZE) is an ASR system that lets the user conduct speaker comparisons using various features and algorithms flexibly. The study in [63] presented a new DNN-based version of VOCALIZE using x-vector, which offered a choice of feature extraction, speaker modeling, and speaker comparison approaches and allowed for speech recording at various training stages. In their work, they used landline recordings and mobile recordings (GBR-ENG) for comparison purposes. The results revealed that the adapted EER for the x-vector for mobile was 1.40%, while the adapted EER for the i-vector was 5.80%. The x-vectors outperformed the i-vectors, especially for short durations.

4.7. Features combination

The study by [64] proved that the use of phase information extraction, which is often ignored in the conventional speaker recognition method based on MFCCs, could improve the speaker IR from 97.4% to 98.8%. GMM was used for modeling the speaker model, and two sets of databases were used in this research to evaluate their method: the NTT and the Japanese Newspaper Articles Sentences (JNAS).

The research conducted by [65] proposed several enhancements for the deep neural network (DNN) feature learning. They presented a phone dependent DNN model to supply phonetic information when learning speaker features and proposed two scoring methods: segment pooling and dynamic time warping (DTW). The i-vectors and d-vectors were the two kinds of speaker vectors used in their research. The former is based on the Gaussian model, while the latter is based on neural networks. Their database comprised 100 speakers who uttered ten short phrases that contained 2–6 Chinese characters. They managed to combine the best i-vector system probabilistic linear discriminant analysis (PLDA) and the best d-vector strategy (DNN+DTW) to give a ~2% Equal Error Rate (EER). It was proven that the combination led to the best performance.

In preventing unauthorised persons from directly or indirectly attacking the speaker recognition system, [66] proposed a new security level using watermark technology. Their method hid the data by changing the sample's frame size in the speech signal's unvoiced part to prevent other people from noticing such data. Both the MFCCs and the Linear Prediction Coding (LPC) were used as features. Pearson's Correlation was used as the classifier to investigate whether the speaker can be recognized (authorized) or unrecognized (unauthorised). A database of 10 speakers uttering the same word was recorded. The results showed 100% accuracy in security could be achieved using the watermark technology along with 93.33% recognition accuracy. Hence, the watermark technology could be applied to make the system more secure.

ASR via wireless communication is known to cause degradation because of synthesized speech. The research work in [67] proposed using the Gaussian probabilistic linear discriminant analysis (GPLDA) as the classifier and the feature selection using Linear Discriminant Analysis (LDA) and the result rejects 19 MFCCs features out of 69 features to improve the performance accuracy. TIMIT_8, the modified version of the TIMIT corpus, was used with 130 speakers. The results showed that the EER for the uncoded speech was 0.91%, while that the EER for the synthesized speech was 12.5%.

The work by [68] proposed a novel fusion of MFCC and time-based (MFCCT) features that were fed as input to the DNN to construct the speaker identification model. LibriSpeech dataset which consists of 100 speakers was used. Besides DNN, five other machine learning classification algorithms (i.e., Random Forest, k Nearest Neighbors, Naïve Bayes, J48, and Support Vector Machine) were used for comparison purposes. The results revealed that DNN outperformed the other five classification algorithms with an overall accuracy of 92.9%.

The research in [69] aimed to identify the speakers under clean and noise background using a limited dataset. They proposed a multitaper based on MFCC and power normalization cepstral coefficients (PNCC) as feature vectors. These features were then normalized using cepstral mean and variance normalization (CMVN) and feature warping (FW). TIMIT and SITW 2016 were the databases used for evaluation. The results indicated that their proposed method provided better performance accuracy compared to the other state-of-art techniques. The accuracy using the TIMIT database was 92.52%, 86.7%, 85.7%, and 85.96% under clean speech, AWGN, babble, and street, respectively.

Table 6 summarizes the progress of research in the area of speaker recognition in the past decade.

Conclusion

In this survey, the aim is to explore speaker recognition in depth, starting by describing the types of biometrics and their characteristics. The difference between speaker recognition and speech recognition was highlighted in this paper to avoid misusing these words. Speaker recognition and its applications in the real-world were explained. A summary of the chronology of advancements in speaker recognition in the past decade was presented. The technology of speech signal processing was explained, and issues related to variability, insufficient data, and background noise were identified as challenges in getting robust speaker recognition systems.

Table 6
Progress in speaker recognition in the last decade.

Author/ Year	Features extraction	Method	Database(Language)	Population(No. of speakers)	Accuracy (%)
[49]/ 2010	DWT (Daubechis, Symlets, Coiflets)	MLP	Self-generated - 2 Czech Speaker Corpora (Czech language)	Corpora 1 – 10, Corpora 2 - 50	SYML20 with MLP = 98% IR
[50]/ 2011	MFCCs	FMMNN	Self-generated (Marathi Language)	50	99.9% IR
[51]/ 2011	MFCCs	GMM-UBM	TIMIT (English Language)	100	80% for both limited and noisy data
[52]/ 2011	13 MFCCs + 13Δ + 13ΔΔ	CHMM	Self-generated (Arabic Language)	10	100% for text dependent, 80% for text-independent
[53]/ 2012	13 MFCCs + 13Δ + 13ΔΔ	FCM + FSVM	KING Speech (English Language)	51	98.76% IR
[54]/ 2013	MFCCs	HMM+GFM (fusion)	VoxForge, NIST 2003 (English Language)	100, 140	93% IR
[55]/ 2014	MFCCs	LFA-SVM Linear Kernel	TIMIT (English Language)	38	82.84% IR
[56]/ 2017	MFCCs	NN	ELSDSR (English Language)	22	93.2%
[57]/ 2019	MFCCs	CNN	Self-generated (English, Hindi & Marathi Language), SRL82 (Chinese Language)	50	CNN = 71% IR for SRL82; CNN = 75% IR for real-world voice sample
[58]/ 2020	MFCCs	AFEASI	LibriSpeech (English language)	251	AFEASI = 0.95 accuracy
[59]/ 2013	TTESBCC	GMM	Self-generated (Marathi Language)	25	Neutral speech = 98.6% IR; Whisper speech = 55.8% IR
[60]/ 2015	NDSF	Not mentioned	Self-generated (English Language), MVSR-IITG (Hindi Language)	100	98%−100% IR
[61]/ 2018	Short-term magnitude spectrogram	CNN (ResNet)	VoxCeleb2, VoxCeleb1 (Multi-languages)	6000 1251	ResNet-50 = 3.95% EER
[62]/ 2019	x-vectors	E-TDNN	NIST SRE18: CMN2 (Arabic language), VAST	~4500 ~7000	E-TDDN = 4.95% EER; E-TDDN = 11.1% EER
[63]/ 2019	x-vectors	DNN (VOCALIZE)	Mobile recordings (GBR-ENG), Landline recordings (English language)	534 387	x-vector = 1.40% EER
[64]/ 2012	MFCCs + Phase Information	GMM	NTT, JNAS (Japanese Language)	35, 270	98.8% IR
[65]/ 2015	i-vector + d-vector	PLDA, DNN+DTW	Self-generated (Chinese Language)	100	~ 2% EER
[66]/ 2016	MFCCs + LPCs	Pearson's correlation	Self-generated (not mentioned)	10	WRA = 100%, 93.33% IR
[67]/ 2018	LDA+MFCCs	GPLDA	TIMIT_8 (English Language)	130	Uncoded speech = 0.91% EER, Synthesized speech = 12.5% EER
[68]/ 2020	MFCCT (MFCC + time-based)	DNN	LibriSpeech (English language)	100	92.9%
[69]/ 2020	Multitaper (MFCC + PNCC)	ELM	TIMIT	124	97.52% for clean speech, 86.70% for AWGN noise, 85.70% for babble noise, 85.96% for street noise

Adversarial attacks were also discussed as they have become a serious issue when dealing with machine learning and deep learning. The structure of speaker recognition and the choices on classifiers were explained thoroughly.

There is currently a great demand for the development of technologies that integrate biometric systems due to their wide range of applications, especially when the identification of the individual is needed. Following years of research and development, devices that use voice as the primary mode of interaction, such as Google Home, Amazon Echo (Alexa), Apple's Siri, and Samsung's Bixby, are now

widely available. Most of these are used in households where multiple people are expected to interact with the device. Such tools and technology gain popularity not only for home usage but also for making the interaction between human and humanoid robots more realistic. Despite the vast amount of work in this area, there are still significant challenges in getting highly accurate systems for practical scenarios.

Declaration of Competing Interest

None.

Acknowledgment

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research.

References

[1] Biometrics: Authentication & Identification – 2020 Review. (2019). Retrieved from https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/biometrics.
[2] Zheng TF, Li L. Robustness-related issues in speaker recognition. Springer; 2017.
[3] De Luis-García R, Alberola-López C, Aghzout O, Ruiz-Alzola J. Biometric identification systems. Signal Process 2003;83:2539–57. https://doi.org/10.1016/j.sigpro.2003.08.001.
[4] Singh N. A study on speech and speaker recognition technology and its challenges. In: Proceedings of national conference on information security challenges; 2014. p. 34–6.
[5] Sharma, A.M. (2019). Speaker recognition using machine learning techniques. (Master's Projects). Retrieved from https://scholarworks.sjsu.edu/etd_projects/685.
[6] Zhang Z. Mechanics of Human Voice Production and Control. J Acoust Soc Am 2016;140(4):2614. https://doi.org/10.1121/1.4964509.
[7] Furui S. 40 years of progress in automatic speaker recognition. In: Tistarelli M, Nixon MS, editors. Advances in biometrics". ICB 2009. lecture notes in computer science, 5558. Berlin, Heidelberg: Springer; 2009.
[8] Sujiya S, Chandra E. A review on speaker recognition. Int J Eng Technol (IJET) 2017;9(3):1592–8.
[9] Imam SA, Bansal P, Singh V. Review: speaker recognition using automated systems. AGU Int J Eng Technol (AGUIJET) 2017;5:31–9.
[10] Kershet J, Bengio S. Introduction. In: Margin L, Methods K, editors. Automatic speech and speaker recognition. West Sussex, United Kingdom: John Wiley & Sons Ltd; 2009.
[11] Kikel, C. (2019, November 11). Difference between voice recognition and speech recognition. Retrieved from https://www.totalvoicetech.com/difference-between-voice-recognition-and-speech-recognition/.
[12] Hanifa RM, Isa K, Mohamad S. Malay speech recognition for different ethnic speakers: an exploratory study. In: 2017 IEEE symposium on computer applications & industrial electronics (ISCAIE); 2017. p. 91–6.
[13] Biometric Today. (2018). Retrieved from https://biometrictoday.com /5-differences-between-voice-and speech-recognition/.
[14] Sharma V, Bansal PK. A review on speaker recognition approaches and challenges. Int J Eng Res Technol 2013;2(5):1580–8.
[15] Kaphungkui NK, Kandali AB. Text dependent speaker recognition with back propagation neural network. Int J Eng Adv Technol (IJEAT) 2019;8(5):1431–4.
[16] Suchitha TR, Bindu AT. Feature extraction using MFCC and classification using GMM. Int J Sci Res Dev (IJSRD) 2015;3(5):1278–83.
[17] Singh N, Agrawal A, Ahmad Khan R. A critical review on automatic speaker recognition. Sci J Circ, Syst Signal Process 2015;4(2):14–7.
[18] Ibrahim, Y.A., Odiketa, J.C. & Ibiyemi, T.S. (2017). Preprocessing technique in automatic speech recognition for human computer interaction: an overview. Retrieved from https://anale-informatica.tibiscus.ro/download/lucrari/15-1-23-Ibrahim.pdf.
[19] Cutajar M, Gatt E, Grech I, Casha O, Micallef J. Comparative study of automatic speech recognition techniques. IET Signal Proc 2013;7(1):25–46.
[20] Malik S, Afsar FA. Wavelet transform based automatic speaker recognition. In: IEEE 13th international multitopic conference; 2009. p. 1–4.
[21] Singh N, Khan RA, Shree R. Applications of speaker recognition. Procedia Eng 2012;38(2012):3122–6. https://doi.org/10.1016/j.proeng.2012.06.363. ISSN 1877-7058.
[22] Zheng T, Li L. Robustness-related issues in speaker recognition. Springer; 2017.
[23] Gupta H, Gupta D. LPC and LPCC method of feature extraction in speech recognition system. In: *2016 6th international conference - cloud system and big data engineering (Confluence),* Noida, 2016; 2016. p. 498–502. https://doi.org/10.1109/CONFLUENCE.2016.7508171.
[24] Kaur K, Jain N. Feature extraction and classification for automatic speaker recognition system – a review. Int J Adv Res Comput Sci Softw Eng 2015;5(1):1–6.
[25] Jamal N, Shanta S, Mahmud F, Sha'abani MNAH. Automatic speech recognition (ASR) based approach for speech therapy of aphastic patients: a review. In: AIP Conference Proceedings. 1883; 2017.
[26] Kumar P, Chandra M. Hybrid of wavelet and MFCC features for speaker verification. In: IEEE world congress on information and communication technologies (WICT); 2011. p. 1150–4.
[27] Mohammed, R.A., Ali, A.E. & Hassan, N.F. (2019). Journal of Al-Qadisiyah for computer science and mathematics, 11(3), pp. 21–30.
[28] Janse PV, Magre SB, Kurzekar PK, Deshmukh RR. A comparative study between MFCC and DWT feature extraction technique. Int J Eng Res Technol (IJERT) 2014;3(1):3124–7.
[29] Gomar MG. System and method for speaker recognition on mobile devices. Google Patents; 2015.
[30] Nematollahi MA, Al-Haddad SAR. Distant speaker recognition: an overview. Int J Humanoid Rob 2015;12:1–45.
[31] Rosenberg, et al. L16: speaker recognition. Lect Slides 2007. Retrieved from, http://research.cs.tamu.edu/prism/lectures/sp/l16.pdf.
[32] Gish H, Schmidt M. Text-independent speaker identification. Signal Process Mag, IEEE 1994;11:18–32.
[33] Gaikward SK, Gawali BW, Yannawar P. A review on speech recognition technique. Int J Comput Appl 2010;10(3):16–24.
[34] Sturim, D.E., Campbell, W.M. & Reynolds, D.A. (2007). Classification methods for speaker recognition. proceedings of speaker classification i: fundamentals, features, and methods, pp. 278–97. 10.1007/978-3-540-74200-5_16.
[35] Li Z, Li L. Robustness-related issues in speaker recognition. Springer Briefs Signal Process 2017:39–48.
[36] Muda L, Begam M, Elamvazuthi L. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. J Comput 2010;2(3):138–43.
[37] Tolba H. A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach. Alexandr EngJ 2011 2011;50:43–7. https://doi.org/10.1016/j.aej.2011.01.007.
[38] Nakagawa S, Wang L, Ohtsuka S. Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio, Speech, Lang Process 2012;20:1085–95.
[39] Sandouk, U. (2012). Speaker recognition: speaker diarization & identification. (Master's Thesis). Retrieved from https://studentnet.cs.manchester.ac.uk/resources/library/thesis_abstracts/ProjProgReptsMSc12/Sandouk-Ubai-ProgressReport.pdf.
[40] Doddington GR. Speaker recognition – identifying people by their voices. Proc IEEE 1985;73(11):1651–65.
[41] Kotti M, Moschou V, Kotropoulos C. Speaker segmentation and clustering. Signal Process 2007;88(5):1091–124.

[42] Haohui (2019). Adversarial attacks in machine learning and how to defend against them. Notes from the keynote speaker speech by Professor Ling Liu at the 2019 IEEE big data conference. Retrieved from: https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them-a2beed95f49c.
[43] Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE, symposium on security and privacy workshops; 2018.
[44] Yuan X, Chen Y, Zhao Y, Long Y, Liu X, Chen K, Zhang S, Huang H, Way X, Gunter CA. CommanderSong: a systematic approach for practical adversarial voice recognition. In: Proceedings of the 27th USENIX security symposium, August 15 – 17; 2018. p. 49–64. ISBN: 978-1-939133-04-5.
[45] Li Z, Shi C, Xie Y, Liu J, Yuan B, Chen Y. Practical adversarial attacks against speaker recognition system. In: Proceedings of the 21st international workshop on mobile computing systems and applications (HotMobile '20), March 3–4; 2020. https://doi.org/10.1145/3376897.3377856. Austin, TX, USA. ACM, New York, NY, USA, 6 pages.
[46] Singh N, Agrawal A, Khan RA. The development of speaker recognition technology. Int J Adv Res Eng Technol (IJARET) 2018;9(3):8–16.
[47] Sharma A, Singla SK. State-of-the-art modeling techniques in speaker recognition. Int J Electron Eng 2017;9(2):186–95.
[48] Shaver CD, Acken JM. A brief review of speaker recognition technology. Electr Comput Eng Fac Publ Presentat 2016;350. Retrieved from, https://pdxscholar.library.pdx.edu/ece_fac/350.
[49] Král P. Discrete wavelet transform for automatic speaker recognition. In: Image and signal processing (CISP), 2010 3rd international congress on, 2010; 2010. p. 3514–8.
[50] Jawarkar N, Holambe R, Basu T. Use of fuzzy min-max neural network for speaker identification. In: Recent trends in information technology (ICRTIT), 2011 International Conference on; 2011. p. 178–82.
[51] Krishnamoorthy P, Jayanna HS, Prasanna SRM. Speaker recognition under limited data condition by noise addition. Expert systems with applications, 38. Elsevier. 2011; 2011. p. 13487. https://doi.org/10.1016/j.eswa.2011.04.069.
[52] Tolba H. A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach. Alex Eng J 2011 2011;50:43–7. https://doi.org/10.1016/j.aej.2011.01.007.
[53] YuJuan X, Hengjie L, Ping T. Hierarchical fuzzy speaker identification based on FCM and FSVM. In: Fuzzy systems and knowledge discovery (FSKD), 2012 9th international conference; 2012. p. 311–5.
[54] Bhardwaj S, Srivastava S, Hanmandlu M, Gupta J. GFM-based methods for speaker identification. IEEE Trans Cybern 2013;43:1047–58. 2013.
[55] Shen X, Zhai Y, Wang Y, Chen H. A speaker recognition algorithm based on factor analysis. In: 2014 7th international congress on image and signal processing; 2014. p. 897–901.
[56] Soleymanpour M, Marvi H. Text-independent speaker identification based on selection of the most similar feature vectors. Int J Speech Technol 2017;20:99–108. 2017.
[57] Jagiasi R, Ghosalkar S, Kulal P, Bharambe A. CNN based speaker recognition in language and text-independent small scale system. In: Proceedings of 3rd international conference on IoT in social, mobile, analytics and cloud (I-SMAC); 2019. p. 176–9.
[58] Li R, Jiang J-Y, Liu J, Hsieh C-C, Wang W. Automatic speaker recognition with limited data. In: the 13th ACM international conference on web search and data mining (WSDM '20); 2020.
[59] Jawarkar NP, Holambe RS, Basu TK. Speaker identification using whispered speech. In: Communication systems and network technologies (CSNT), 2013 international conference on, 2013; 2013. p. 778–81.
[60] Chougule SV, Chavan MS. Robust spectral features for automatic speaker recognition in mismatch condition. Procedia Comput Sci 2015;58:272–9. https://doi.org/10.1016/j.procs.2015.08.021.
[61] Chung, J.S., Nagrani, A. & Zisserman, A. (2018). VoxCeleb2: deep speaker recognition. arXiv:1806.05622.
[62] Villalba J, Chen N, Snyder D, Garcia-Romero D, McCree A, Sell G, Borgstrom J, Richardson F, Shon S, Grondin F, Dehak R, Carcia-Perera LP, Povey D, Torre-Carraswuillo PA, Khudanpur S, Dehak N. State-of-the-art speaker recognition for telephone and video speech: the JHU-MIT submission for NIST SRE18. Proc. Interspeech 2019 2019:1488–92. https://doi.org/10.21437/Interspeech.2019-2713.
[63] Kelly F, Forth O, Kent S, Gerlach L, Alexander A. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In: AES international conference on audio forensics; 2019.
[64] Nakagawa S, Wang L, Ohtsuka S. Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio, Speech, Lang Process 2012;20:1085–95.
[65] Li L, Lin Y, Zhang Z, L, Wang D. Improved deep speaker feature learning for text-dependent speaker recognition. In: APSIPA annual summit and conference; 2015. p. 426–9.
[66] Desai N, Tahilramani N. Digital speech watermarking for authenticity of speaker in speaker recognition system. In: International conference on micro-electronics and telecommunication engineering; 2016. p. 105–9.
[67] Zergat KY, Selouani SA, Amrouche A. Feature selection applied to G.729 synthesized speech for automatic speaker recognition. In: IEEE 5th international congress on information science & technology (CIST); 2018. p. 178–82.
[68] Jahangir R, Teh YW, Memon NA, Mujtaba G, Zareei M, Ishtiaq U, Akhtar MZ, Ali I. Text-independent speaker identification through feature fusion and deep neural network. IEEE Access 2020;8:32187–202. https://doi.org/10.1109/ACCESS.2020.2973541.
[69] Bharat KP, Rajesh KM. ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score. Multimed Tools Appl 2020;79:28859–83. https://doi.org/10.1007/s11042-020-09353-z.

Rafizah Mohd Hanifa obtained her bachelor's degree in computer science from Universiti Sains Malaysia (USM) in 1999, followed by a master's degree in Information Technology from Universiti Utara Malaysia (UUM) in 2001. She is currently a Ph.D. student at Universiti Tun Hussein Onn Malaysia (UTHM). Her research interests include speech processing, artificial intelligence, and augmented reality.

Khalid Isa graduated from Universiti Teknologi Malaysia in 2001 with a BSc in Computer Science. He pursued his MSc. in Computer Systems Engineering and Communications at Universiti Putra Malaysia, graduated in 2005. In 2014, he completed his Ph.D. degree in Electrical and Electronic Engineering at Universiti Sains Malaysia, specialized in Computational Intelligence and Underwater Robotics.

Shamsul Mohamad obtained his BSc and MSc in Computer Science from Universiti Teknologi Malaysia in 1999 and Universiti Sains Malaysia in 2004 respectively. He completed his Ph.D. degree in Computer Science at Universiti Teknologi Malaysia, with a specialization in Crowd Simulation. His-research interests include crowd simulation, artificial intelligence, and the Internet of Things.