Multimedia Tools and Applications (2021) 80:32725–32762
https://doi.org/10.1007/s11042-021-11235-x

A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction

Choon Beng Tan[1] · Mohd Hanafi Ahmad Hijazi[1] · Norazlina Khamis[1] · Puteri Nor Ellyza binti Nohuddin[2] · Zuraini Zainol[3] · Frans Coenen[4] · Abdullah Gani[1]

Received: 25 June 2020 / Revised: 2 July 2021 / Accepted: 8 July 2021 /
Published online: 4 August 2021

Abstract

The emergence of biometric technology provides enhanced security compared to the traditional identification and authentication techniques that were less efficient and secure. Despite the advantages brought by biometric technology, the existing biometric systems such as Automatic Speaker Verification (ASV) systems are weak against presentation attacks. A presentation attack is a spoofing attack launched to subvert an ASV system to gain access to the system. Though numerous Presentation Attack Detection (PAD) systems were reported in the literature, a systematic survey that describes the current state of research and application is unavailable. This paper presents a systematic analysis of the state-of-the-art voice PAD systems to promote further advancement in this area. The objectives of this paper are two folds: (i) to understand the nature of recent work on PAD systems, and (ii) to identify areas that require additional research. From the survey, a taxonomy of voice PAD and the trend analysis of recent work on PAD systems were built and presented, whereby the recent and relevant articles including articles from Interspeech and ICASSP Conferences, mostly indexed by Scopus, published between 2015 and 2021 were considered. A total of 172 articles were surveyed in this work. The findings of this survey present the limitation of recent works, which include spoof-type dependent PAD. Consequently, the future direction of work on voice PAD for interested researchers is established. The findings of this survey present the limitation of recent works, which include spoof-type dependent PAD. Consequently, the future direction of work on voice PAD for interested researchers is established.

Keywords Speaker identification · Speaker verification · Anti-spoofing voice recognition · Voice presentation attack detection · Voice PAD

✉ Mohd Hanafi Ahmad Hijazi
hanafi@ums.edu.my

[1] Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, Sabah, Malaysia

[2] Institute of IR4.0, Universiti Kebangsaan Malaysia, Selangor, Malaysia

[3] Department of Computer Science, Faculty of Science and Defence Technology, Universiti Pertahan Nasional Malaysia, Kuala Lumpur, Malaysia

[4] Department of Computer Science, University of Liverpool, Liverpool, UK

1 Introduction

Biometric is a process of identifying and differentiating between individuals based on the differences in biological and behavioral characteristics. According to the National Science & Technology Council's (NSTC) Subcommittee on Biometrics, biometric is common terminology used to narrate a characteristic or a process [11]. When biometric is used to describe a characteristic, it refers to the quantifiable biological and behavioral characteristics which could be used for automated recognition. Likewise, when biometric is used to narrate a process, it refers to the methods of automatically recognizing a biometric subject based on observable biological and behavioral properties.

Biometric can be categorized into two main types, namely physiological and behavioral biometrics [100]. Physiological biometrics refers to the distinct characteristics that are related to an individual's physical body shape like DNA, eyes (iris and retina), fingerprint, and face [12]. On the other hand, behavioral biometrics refers to the unique characteristics that are related to an individual's behavioral patterns like typing rhythm, voice, and human motion. Examples of biometric technologies that have been applied widely in societies are fingerprint recognition-based immigration control, virtual assistant via speech recognition, and smartphone login using face recognition.

Voice biometric can be applied in many ways. For example, it can be used in healthcare for voice disorder detection, assists in voice disorder assessment and treatment [81]. An application of voice biometric, namely voice recognition or speaker recognition refers to the process of recognizing the person who is speaking. Voice recognition can be further classified into two categories, namely speaker identification [133] and speaker verification [24]. Speaker identification refers to the process of identifying the speaking person, whereas speaker verification refers to the process of verifying the claimed identity of the speaking individual, as presented in Fig. 1. Voice recognition uses both physiological and behavioral components in identifying and verifying the identity of the speaker. Some applications of voice recognition are access control, forensic criminal investigation, surveillance of phone conversation, and banking transaction [68].

Despite the benefits brought by voice recognition technology, spoofing attacks from security adversaries is inevitable. A spoofing attack refers to a malicious party launching an attack to impersonate an authorized individual in the voice recognition system to bypass and get access to the system. Due to the ease of obtaining biometric data via social media such as Facebook, Instagram, and WhatsApp [71], countermeasures against spoofing attacks are needed to enhance the security of biometric systems. These countermeasures

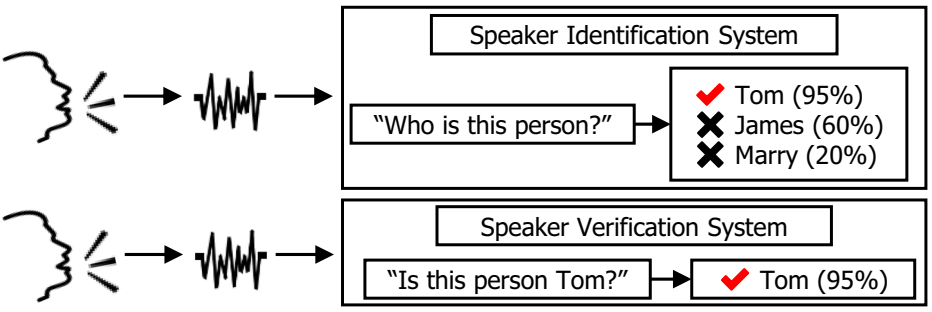


Fig. 1 Illustration of speaker identification versus speaker verification

are known as voice Presentation Attack Detection (PAD). However, progress in PAD in the field of speaker recognition does not receive equal attention as other types of biometric such as fingerprint and face recognition [28]. Some of the reasons that affect the progress in anti-spoofing measures in the speaker recognition field are late invention, deployment, and limited applications of voice recognition technology in the past compared to biometrics like fingerprint and face recognition [73]. Nevertheless, the hands-free property of voice recognition has made it widely accepted and applied to various fields, not limited only to smartphone login, voice verified bank transaction, and access control verification [69]. Hence, an effective ready-to-use voice PAD system for voice recognition is required as there were no publicly available finished products of these voice PAD systems [114].

There are numerous works recently conducted on voice PAD that can be found in the literature. However, to the best of our knowledge, four articles [49, 88, 103, 136] presented a survey and indexed in Scopus. The most similar [103] was published in 2019, that present reviews and summarizes some voice PAD for speaker recognition systems. Article [88] focuses only on replay attacks while two articles [49, 143] focus only on the voice PAD presented in ASVspoof Challenges. Meanwhile, the article [103] published in 2019, presented all four types of presentation attacks. However, most of the papers did not provide a descriptive taxonomy on recent voice PADs. A taxonomy categorizes previous work based on the identified attributes that could help readers understand the topic better. Hence, the survey presented in this paper aims to expand the domain of knowledge by providing the categorization of the related work and building taxonomy from the most recent work on voice PAD, which includes those presented in the ASVspoof2019 Challenge. Besides, this paper also contributed by providing the trends and analyses of voice PAD, which are lacking in the other survey articles. The issues and future direction of voice PAD are also described in this paper.

The paper has contributions:

- To produce a taxonomy on recent voice PAD systems.
- To visualize trends of work on PAD.
- To identify the issues faced by current voice PAD and describes corresponding future directions of PAD.

The remaining of this survey paper is arranged according to the following. Section 2 described the methodology used to conduct this survey. The recent speaker verification systems published in the last five years are presented in Section 3. The findings of the survey, which include voice spoofing attacks and PAD, analysis on the trend of recent voice PAD, research gaps, and future direction of voice PAD systems are presented in Section 4 This survey paper is concluded in Section 5.

2 Methodology

In this section, the methodology used to survey recent speaker verification systems, voice spoofing, and voice PAD are described.

First, to identify the recent work on speaker verification systems, we referred to a variety of sources, including online resources such as news, forums, scientific materials that include journals and conference articles. Online resources are used to retrieve up-to-date information about the applications of speaker recognition, as well as speaker recognition security risks, issues, and incidents which have been identified or happened. Meanwhile, scientific

materials such as journals and conference articles are included in this survey to assess the state-of-the-art speaker recognition systems and corresponding types of voice PAD to secure speaker recognition systems.

To assure that this survey only covers the state-of-the-art speaker verification systems and voice PAD, only related scientific materials published in recent years (2015–2021) are considered. Nonetheless, several older but significant articles are included, as well. A total of 172 Scopus indexed articles are considered in this survey. Recent articles presented in Interspeech and ICASSP Conferences are also included in this paper. To search for all possible voice PAD articles indexed in Scopus, common keywords such as "voice" and "anti-spoofing" are used to search for the PAD articles. As technical terms such as "presentation attack detection" and "PAD" may miss out on some relevant articles, hence these technical terms are not selected as keywords. The most recent survey paper on voice PAD was published in 2019 [103].

3 Speaker Verification

Voice recognition, commonly known as speaker recognition, refers to recognizing the speaking person, whereas speech recognition refers to recognizing the words from speech. Voice recognition is grouped into two categories, namely speaker identification and speaker verification, as illustrated in Fig. 1. Voice recognition uses both physiological and behavioral components in identifying and verifying the identity of the speaker.

Speaker verification is a process where the claimed identity of the owner of the voice, the target voice, is verified by comparing the target voice with the registered voice in the database [47]. Hence, speaker verification is a 1:1 matching process between the target voice and voices registered in the database [4]. The main application of speaker verification is on authentication such as verification of identity in phone banking transactions and voice authenticated access control for door lock [69, 109].

There are two phases in Automatic Speaker Verification (ASV) systems, namely the speaker enrollment phase and speaker verification phase (speaker verification phase) [117]. In the speaker enrollment phase, the aim is to generate the speaker models. First, features are extracted from the voice captured by the voice recognition system. Second, the features are used to generate the speaker model. Last, the generated speaker model is enrolled in the database of the voice recognition system. This process is drawn and shown in Fig. 2. The verification of the speaker is conducted in the speaker verification phase. First, features are extracted from the voice by the voice recognition system. At the same time, the claimed or targeted speaker model is retrieved from the database. Second, the patterns of extracted features are matched with the retrieved speaker model. If the matching obtained a score equal to or greater than the threshold set in the voice recognition system, then the claim of the identity by the speaker is accepted; otherwise, the claim is rejected. Fig. 3 was drawn to show the process of speaker verification.

Based on Figs. 2 and 3, the three key components in speaker verification systems include feature extraction, speaker modeling, and pattern matching. During pattern matching, the speaker verification system will either accept or reject the claim of identity based on the score of pattern matching [94]. Equal Error Rate (EER) is often the performance measures used for speaker verification [36, 87, 95, 113, 143]], a metric that is commonly used to assess the biometric system performance. EER represents the value of error in which the error rates, namely the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are

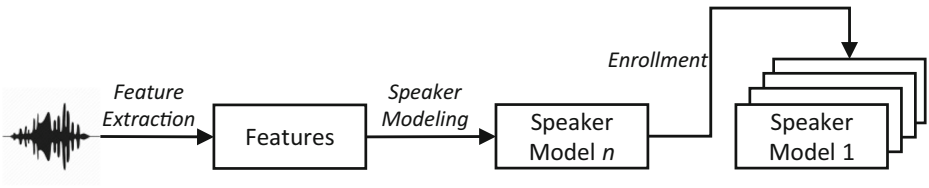


Fig. 2 Speaker enrollment phase

equal. Note that FAR refers to the probability of incorrectly accepts an unauthorized access attempt by a biometric system. In contrast, FRR refers to the probability of wrongly rejects an authorized access attempt by a biometric system. The better the performance of speaker verification, the lower the EER is [95].

From the literature, the Gaussian Mixture Model (GMM) [95–97] is the method that produced a more robust and better-performing speaker verification system than other speaker modeling approaches. Therefore, it has been extensively used for feature extraction for speaker verification in recent works [1, 107, 115]. GMM is a probabilistic model describing normally distributed subpopulations within an overall population and was used in voice recognition feature extraction. To verify the identified speaker from a speech, GMM compares the captured voice with a general, person-independent speaker model. The Universal Background Model (UBM) [46, 67, 76] is often being used as the general model for GMM. In speaker verification, UBM is a general model used to represent general feature characteristics that can be used to compare against the specific person being verified. Researchers also have successfully applied other speaker modeling techniques such as i-vector [21, 36], and x-vectors [116] for front-end feature extraction. i-vectors was introduced as a simple model for speaker recognition in which the feature extraction was conducted using simple factor analysis. x-vectors were the fixed-dimensional embeddings extracted with DNN for speaker recognition, and it was found that the x-vector-based system out-performed the standard i-vector-based system.

On the other hand, back-end classifiers such as Deep Neural Network (DNN) [2, 24, 115] and Probabilistic Linear Discriminant Analysis (PLDA) [121] were shown to be able to discriminate between spoof and genuine speech signals with low EER using features like i-vectors and x-vectors. Recently, there were a number of end-to-end approaches proposed for speaker verification [40, 64]. An end-to-end approach, in the context of speaker verification, is a model or classifier that is trained together with the feature learning. Nonetheless, another approach emerged that focuses on learning speaker features while leaving the classifier as

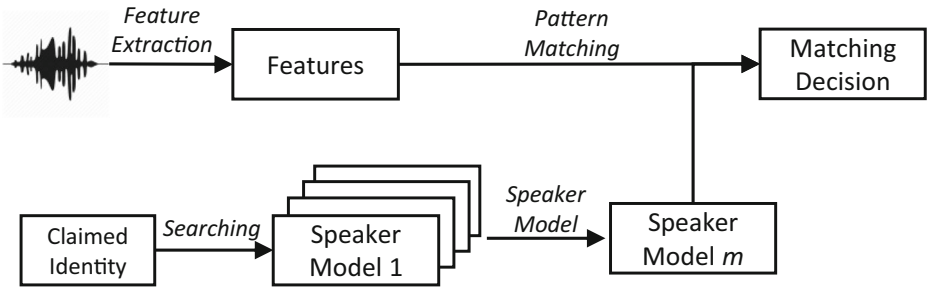


Fig. 3 Speaker verification phase

a separate component. This approach indicates that if the feature learning is strong enough, then the limitation of the classifier will become negligible. Compared to the end-to-end approach, the feature learning approach was found to outperform the end-to-end approach consistently, based on a dataset consisting of utterances from 5,000 speakers [129].

To further improve the performance of the ASV system, some works used score normalization. In [8], it has been shown that score normalization can lead to not only improved performance but also better calibration and a more reliable threshold for ASV systems. For example, an improvement of 30% was produced by the adaptive symmetric score normalization (s-norm) by the work [70] using NIST SRE 2016 dataset. Another method to improve further the performance of a speaker verification system is using fusion [10, 108]. There are two frequently used fusion techniques for speaker verification, namely, score fusion [55] and feature fusion [5]. Score fusion is a method used to make a final decision by matching the scores output from more than one biometric modal. Score fusion can be conducted by combining scores generated by biometric models using approaches like logistic regression. Due to its implementation simplicity, score fusion is the most commonly used fusion method in multibiometric systems [98]. However, recent works [32, 55] have shown that the scored fusion outperformed feature fusion as a fused feature is more complex and may lose its significant traits, which can be used to distinguish accurately between different speaking individual [55].

Although the state-of-the-art ASV systems are capable of verifying the claimed identity of speakers with a low error rate and high accuracy, the systems are prone to presentation attacks due to the ease of obtaining biometric data. Voice PAD was introduced to mitigate the problem of ASV against presentation attacks. The next section presents recent works of voice PAD systems.

4 Voice Presentation Attack Detection, Taxonomy, Research Gap, and Future Direction

This section contains three sub-sections to present voice PAD. Section 4.1 presents voice presentation attacks and PAD. The taxonomy of recent voice PAD is described in Section 4.2. Section 4.3 presents the analysis of the trend on recent voice PAD. Section 4.4 presents the research gap and future direction of voice PAD.

4.1 Voice Presentation Attack Detection (PAD)

Voice spoofing attacks can be grouped into two categories; the sensor level attack and transmission level attack [136]. The transmission attack can be avoided and deflected by having a secure transmission protocol and assistance from security software. By comparison, it is much more difficult to defend against sensor level attacks and requires considerable attention due to the ease of obtaining biometric data as described in Section 1 [71]. This sub-section has thus presented an overview of recent works on the sensor level attack. Commonly, sensor level attack is referred to as presentation attack, where adversary tries to bypass the voice recognition system through spoofed voice input. There are four main categories of voice presentation attacks, namely impersonation, replay, voice conversion, and speech synthesis attacks [136].

The first voice presentation attack, the impersonation or zero-effort imposter is a spoofing technique that requires no assistance from electronic devices. Impersonation is carried out by mimicking a specific person's way of speaking. Impersonation is not an effective

ASV spoofing method [55]. However, there is a case where a non-identical twin reporter from British Broadcasting Corporation (BBC News) had successfully spoof the voice recognition system of HSBC to access the bank account of his twin after mimicking his twin brother's voice [111]. Hence, the threats that impersonation posed to ASV systems must not be underestimated. Meanwhile, the recent work indicated that imitation of speech patterns like fundamental frequency and some key format of a speaker is possible. However, mimicry to replicate all characteristics of a targeted voice seems to be physically impossible [113] due to the uniqueness of the human vocal tract. Moreover, recent studies [74] concluded that mimicry is unable to duplicate natural speech regardless of mimicry training and impersonation skill as it is not a natural act. Another work shows that stable spectral peak features which represent invariant vocal tract characteristics of a speaker could be effective in differentiating genuine and imposter voices [113]. Although impersonation is not effective to spoof most ASV systems [55], the other three types of presentation attacks are major threats [136] because the voiceprint used to spoof ASV systems are originated from the genuine speaker.

The second presentation attack, the replay attack, is the most popular type of spoofing attack as it is the simplest to conduct. As biometric data can be obtained easily through social media, replay attacks can be conducted by anyone using recording devices such as smartphones. The replay attack is more straightforward compared to speech synthesis and voice conversion attacks. The replay attack is more likely to be performed by non-professional adversaries to spoof ASV systems as replaying a pre-recorded audio involves little knowledge of audio signal processing. Several replay attack detectors have been developed for ASV systems. For example, the baseline system in the ASVspoof 2017 Challenge which is based on Constant Q Transform Cepstral Coefficients (CQCC) features with 2-class Gaussian Mixture Model (GMM) classifier was recorded an EER of 30.60% on the evaluation dataset [22]. Nonetheless, using the speech frame selection approach [58], the performance of the CQCC-GMM countermeasure improved to 21.60% EER. Replay attack detectors proposed by researchers using Recurrent Neural Networks (RNN) with Filter Bank (Fbank) features have achieved 9.81% EER [16], whereas the one using Deep Neural Networks (DNN) and Support Vector Machine (SVM) classifiers with CQCC and High-Frequency Cepstral Coefficients (HFCC) features have achieved 11.5% EER [82]. The best replay attack detector system in ASVspoof 2017 Challenge hit an EER of 6.73% on the evaluation dataset was a fusion system that adopted several classifiers and features [60].

The third presentation attack is speech synthesis and voice conversion. Unlike replay attack, spoofing speaker verification system using speech synthesis and voice conversion requires knowledge of signal processing [55], which is mostly conducted by professional adversaries. Speech synthesis attack is one of the effective presentation attacks towards ASV systems where Text-To-Speech (TTS) technology is applied by concatenating available pieces of speech data [105]. Recently several Synthetic Speech Detectors (SSDs) were introduced to protect speaker verification systems from speech synthesis attacks [38, 102]. Since most synthetic speeches were generated using parametric vocoders, SSDs that use phase information [89] for synthetic speech detection has been shown to be effective [23]. As a result, phase-based SSD has become state-of-the-art for detecting synthetic speech [85, 106]. Nonetheless, most of the introduced SSD systems are only effective against parametric vocoders, which use minimum-phase filters for speech synthesis. Thus, phase-based SSD are prone to speech synthesis attack from vocoder which uses mixed-phase filters [23].

Voice conversion attack is performed by converting the voice of a spoof attacker into the voice of the target speaker to cheat the ASV systems. Indicators of converted voice such as the absence of natural speech phase information can be extracted as a feature to detect the converted speech from genuine speech. For instance, features such as cosine normalization and frequency derivative of phase spectrum information can be extracted to detect the converted speech with EER of 6.0% and 2.4% respectively [27]. Other features like Local Binary Pattern (LBP) extracted from images generated from speech signals such as spectrogram [27] can also be used to detect artificial signals such as synthesized and voice-converted speech.

There were several efforts done to foster the development of countermeasure to spoofing of ASV systems. The countermeasures were often known as Presentation Attack Detection (PAD). For example, the building of more public datasets such as the ReMASC dataset that consists of genuine and replayed speech corpus collected in realistic voice-controlled systems' usage scenarios [33]. ReMASC contains recordings from 50 speakers of both genders and of different ages and accents. The recordings were composed of 132 voice commands which were collected in four different environment settings with different levels of noise. The four environments were two indoor with settings of quiet and noisy background, one outdoor, and one moving vehicle scenario. Four different microphones were used in the data collection. As the ReMASC corpus was made up of recordings via a variety of microphones instead of a single microphone, it is well-suited for multi-channel voice PAD research such as [34]. Another major effort from the community of spoofing and anti-spoofing for ASV was the ASVspoof Challenge series. There were a total of three ASVspoof Challenges up-to-date, namely ASVspoof 2015, ASVspoof 2017, and ASVspoof 2019. In general, the ASVspoof Challenge series aims to promote the development of a generalized voice spoofing countermeasure to detect varying and unforeseen spoofing attacks using standardized datasets, protocols, and evaluation metrics.

The first series of ASVspoof challenges, the ASVspoof 2015, was held within the scope of a special session at Interspeech 2015. ASVspoof 2015 dataset consists of genuine, synthesized, and voice-converted utterances in which the utterances were collected from 106 speakers (45 male and 61 female). Spoofed utterances for training and development were generated using three voice conversion and two speech synthesis algorithms. All five algorithms used to generate the spoof utterances of the training and development set were used to generate the spoof utterances in the evaluation set. In addition, an additional five algorithms were used to generate more spoof utterances in the evaluation set, referred to as unknown attacks. EER was used as the primary metric to evaluate the performance of submitted countermeasures. In ASVspoof 2015, there were a total of 16 primary submissions used for ranking in the challenge. In general, most of the submissions achieved low EER, which is less than 1% for known attacks in ASVspoof 2015. The best system submitted to ASVspoof 2015, named System A, has used two features namely Mel-Frequency Cepstral Coefficients (MFCC) and Cochlear Filter Cepstral Coefficients Plus Instantaneous Frequency (CFCCIF), and GMM classifiers with score fusion in detecting the spoof speech with an average EER of 1.211% for known and unknown attacks. In particular, System A has achieved an average EER of 0.408% for known attacks and 2.013% for unknown attacks respectively. A similar trend of having higher EER for unknown attacks than known attacks can be seen in all 16 submissions. This trend can be seen as potential overfitting in the countermeasures proposed. One of the identified reasons for getting higher EER when detecting unknown attacks was the unreliability of the counter-measures in detecting S10 attacks, the only attack that was generated using the waveform concatenation approach. Details of all 16

submissions can be referred to [139]. More details regarding the ASVspoof 2015 Challenge can be found in [137, 138].

Due to the shortcoming of ASVspoof 2015, ASVspoof 2017 was organized to highlight the replay attacks which were excluded during ASVspoof 2017. It was held as a special session at Interspeech 2017. The main objective of ASVspoof 2017 was to assess spoofing attack detection accuracy with 'out in the wild' conditions, to detect replay attacks in particular. The ASVspoof 2017 dataset consists of genuine utterances that were based on the text-dependent RedDots corpus [61], a speech corpus made up of utterances from 49 male and 13 female speakers, whereas the spoof utterances were based on the replayed version of RedDots corpus [52]. Similar to ASVspoof 2015, EER was used as the primary metric to evaluate the performance of submitted countermeasures in ASVspoof 2017. In ASVspoof 2017, there were a total of 49 submissions received for the challenge. The best performing system, System S01, achieved an EER of 6.73%. There were six conditions (C1-C6), ranging from replayed recordings in the condition of background noise that was comparably easier to detect, to high-quality replayed recordings that were difficult to detect. The performance of the countermeasures for condition C6 was consistently the worst. This indicates that the state-of-the-art countermeasures, at that time, were prone to the effects of high-quality replay recordings to spoof the ASV systems. The comparison of the spoof detection rate for ASVspoof 2015 and ASVspoof 2017 suggests that the detection of replay attacks is more difficult than speech synthesis and voice conversion attacks. Hence, the work [51] has highlighted that the generalization of countermeasures remains an open problem. Readers are referred to [22, 50] for more details on the ASVspoof 2017 Challenge.

Similar to the previous editions, the most recent ASVspoof 2019 was held as a special session at Interspeech 2019. The ASVspoof 2019 Challenge extended the previous ASVspoof challenges in several aspects [7]. First, the ASVspoof 2019 covered all three main types of spoofing attacks, namely speech synthesis or text-to-speech (TTS), voice conversion, and replay attacks [131]. Second, the addition of the latest speech synthesis and voice conversion systems with regards to the ASVspoof 2015. Third, a more well-controlled evaluation setup was used for the assessment of replay countermeasures in ASVspoof 2019 compared to ASVspoof 2017. Lastly, the ASVspoof 2019 aligns the countermeasures with the ASV system more closely compared to ASVspoof 2015 and ASVspoof 2017, which were focused on standalone countermeasures. Although the ASVspoof 2019 Challenge was still a standalone spoofing detection task, the adoption of the tandem Decision Cost Function (t-DCF) metric as the primary performance evaluation measure in the challenge will ensure the results obtained reflect the performance of countermeasures on the reliability of ASV systems. The use of EER as the only evaluation metric may not reflect the reliability of the countermeasures in previous ASVspoof challenges [124]. In ASVspoof 2019 Challenge, there were 48 and 50 submissions received for the Logical Access (LA) and Physical Access (PA) scenarios, respectively [124]. Both EER and t-DCF metrics were used in the evaluation. The best performing system for the LA scenario, System T05, has achieved an EER of 0.22% and a t-DCF of 0.0069%. The best performing system for the PA scenario, System T28, has achieved an EER of 0.39% and a t-DCF of 0.0096%. Nonetheless, the majority of countermeasures could not produce an EER of less than 5% in both LA and PA scenarios.

The t-DCF was a new evaluation metric proposed to address two shortcomings of EER. Firstly, EER may not be a reliable performance measure when ASV and spoof countermeasures are combined. Secondly, the metric EER may be biased against user authentication

applications that have high user prior but a low spoofing attack prior, such as telephone banking. Hence, the work [53] proposed to migrate the performance evaluation from spoof countermeasures-centric to ASV-centric with the aid of t-DCF, a newly introduced performance metric. t-DCF is a generalized DCF metric to enable the evaluation of combined ASV and spoof countermeasures. It extended the conventional DCF used in ASV research to scenarios involving spoofing attacks. As there were two detection systems, namely ASV and spoof countermeasures, each with two possible false alarms, four costs were identified, namely the cost of ASV system rejecting a target trial, the cost of ASV system accepting a non-target trial, the cost of countermeasures rejecting a human trial, and the cost of countermeasures accepting a spoof trial. These four costs were used in the calculation of the t-DCF metric. Besides, the work presented in [53] has shown analysis on top-performing countermeasures in ASVspoof 2015 and 2017 with t-DCF focused on spoofing attacks prior. From the result, EER and t-DCF show differences for higher priors, thus some ranking changes can be observed.

Meanwhile, there have been some interesting researches carried out related to biases in model performance resulting from dataset artifacts [19]. The first example is described as the following. Researchers investigated six features extracted from speech for replay attack detection using GMM. Then, the factors that influence the predictions of the GMM models were determined. As a result, researchers uncovered a feature or cue which the models were exploiting; the initial silence frames of zeros present in genuine signals but absent in spoofed signals. The cue was found to make the GMM based spoof detection system to classify incorrectly [17]. Researchers further investigated whether the biases in the model caused by the cue can be resolved by eliminating the initial frames of zeros from the test files. From the experiments, researchers found out that such an approach helped reduced the error rate of the spoof detection systems. Although the vulnerability of ASV systems to spoofing attacks has initiates the development of countermeasures; still, there was no research done on what did countermeasures are learning to discriminate between genuine and spoof speeches. Recently, researchers investigated the local behaviour of a CNN-based replay detection system submitted to ASVspoof 2017 Challenge using the SLIME algorithm [18]. Researchers found out that the model investigated was using the first 400 milliseconds of audio for most of the spoofing instances to make a prediction. This raised an issue of trustworthiness of the detection systems when these systems were shown to exploit cues from the database which are unrelated to the problem for prediction.

4.2 Voice PAD: The Taxonomy

In this section, a taxonomy of the recent work on PAD systems is presented. The taxonomy is built to summarize and provide a clearer picture of the focus and similarities of work on PAD. Seven attributes were selected for inclusion in the taxonomy. These attributes were chosen as they were imminent and can be found in all the articles being surveyed. The seven attributes included in the taxonomy are types of presentation attack, features, classifiers, fusion, methodology, datasets, and evaluation criteria. Each of the attributes was categorized into groups and sub-attributes. For example, the ‘types of presentation attack’ attribute can be grouped into the ‘device assisted attacks’ and ‘attacks that require no device (zero-effort)’. Works that focused on ‘device assisted attacks’ employed either ‘replay’, ‘speech synthesis and voice conversion’, or ‘multiple types of attack’. Sections 4.2.1 - 4.2.7 describe each of these attributes in detail. Figure 4 summarizes the state-of-the-art voice PAD taxonomy.

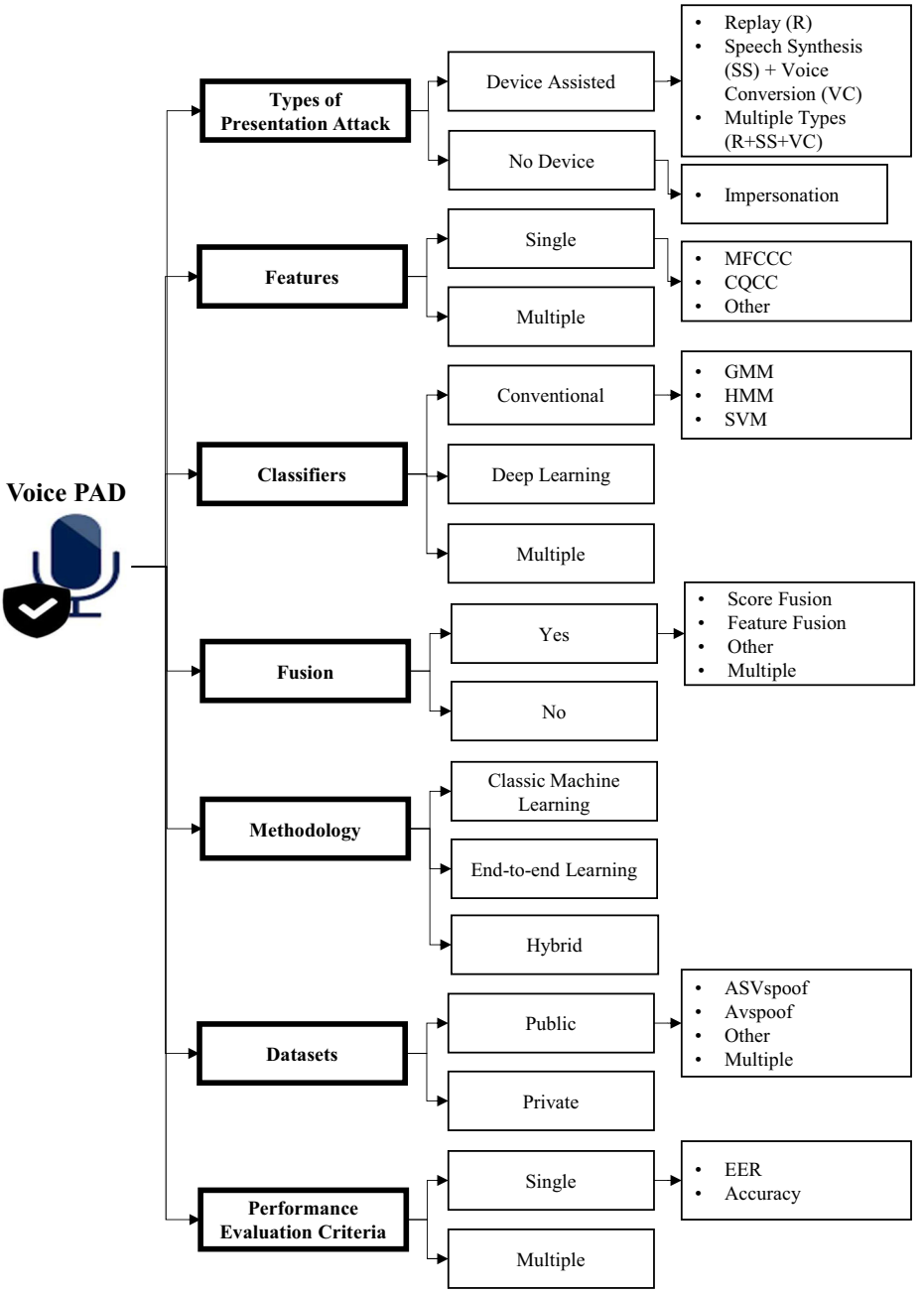


Fig. 4 Summary of voice PAD taxonomy

4.2.1 Types of presentation attack

The first attribute considered in the taxonomy is the types of presentation attacks. The attacks can be grouped into two, (i) the presentation attack using electronic devices and (ii) the presentation attack without an electronic device. Most of the works found were focused on detecting presentation attacks using electronic devices, which were conducted using replay, voice conversion, speech synthesis, or a combination of all.

In the presented taxonomy, speech synthesis and voice conversion were grouped as one subcategory due to these two attacks were similar. They often require the use of an audio processor called vocoder to produce artificial voice. As these attacks require knowledge of on signal processing, assistance from professionals may be needed. Some recent works on speech synthesis and voice conversion detection are [26, 41, 134].

Replay attack has been one of the main focuses in recent works, as it is the most straightforward presentation attack that can be carried out with the aid of electronic devices by the attacker. It can be launched quickly by sneakily records someone's speech and playback the recording to spoof the ASV system. Additionally, unlike attacks on speech synthesis and voice conversion, attackers require no skills and knowledge in signal processing to perform a replay attack. However, replayed speech generation by professional attackers to perform replay attacks do require laborious and time-consuming procedures to yield large databases. Nonetheless, due to the limited availability of replay data, replay detection was not well generalized against unseen conditions, especially channel mismatch conditions [110]. As replay attacks do not require specialized knowledge, the threat of a replay attack can be considered more significant compared to voice conversion and speech synthesis attacks. Some recent works on replay attack detection are [3, 57, 92].

Some works employ a combination of all attacks described above. In these cases, researchers proposed voice PAD systems that counter multiple types of device-assisted presentation attacks. In the actual situation, when an attacker launches a spoofing attack, there is no prior knowledge of the type of attacks being used. As an example, a PAD system developed to counter speech synthesis and voice conversion attacks may ineffective against replay attacks, and vice versa. Recent work showed that a system that is effective against speech synthesis and voice conversion experienced a drastic performance decline when used to differentiate between genuine and replay attacks [142]. Hence, PAD systems to detect spoofing attacks regardless of attack types are needed. Some recent works on PAD capable of detecting multiple types of spoofing attacks are [59, 99, 119, 144].

On the other hand, the impersonation attack or zero-effort imposters was shown to be unable to penetrate most of the state-of-the-art ASV systems [56]. Due to the unavailability of a public dataset for impersonation attacks, there was barely any researches conducted on detecting impersonation attacks on ASV systems. A work, [68] described in [56] experimented on the efficacy of impersonation, in which it turns out that professional imitators are unable to pass the ASV system authentication. However, another research showed the opposite. There was a recent work proposed on detecting speech impersonation [83]. As no public impersonation dataset available for the task, high-quality impersonation speech data was collected. The impersonation corpus was made up of two speakers, 40 genuine and 28 spoof samples. The work used MFCC as the feature and CNN as the classifier for impersonation detection and recorded an EER of 35.85%. The result shown indicates a need to develop a robust countermeasure against impersonation as professional impersonators may succeed in spoofing the ASV system.

4.2.2 Features

The second attribute presented in the taxonomy is features. Some works on PAD used a single feature, while others used more than one. In general, from the surveyed articles, more researchers used multiple features than that of a single feature.

Most single feature PAD systems were based on, but not limited to, MFCC, CQCC, and Linear Predictive Coding (LPC). One of the popular features, MFCC, is the coefficients that make up a Mel-Frequency Cepstrum (MFC) collectively [13]. MFCC was found to be useful for speech synthesis and voice conversion detection [93], though it performed poorly in replay detection [132]. Another popular feature used for voice PAD is the LPC, which is often used as audio features in speech recognition and speaker recognition. In particular, LPC is a technique used frequently in signal processing, in which linear predictive model information is used to represent the spectral envelope of the compressed speech signal [126]. The other popular feature used in PAD, CQCC, is a coefficient extracted from Constant Q Transform (CQT). Recent work has shown that the application of CQCC in PAD outperformed that of MFCC [123]. It is also shown that CQCC was one of the best performing features for voice spoofing detection [122].

As for multiple features-based PAD systems, popular features are including but are not limited to MFCC, CQCC, and Inverted Mel Frequency Cepstral Coefficients (IMFCC). whereas IMFCC is the characteristics property of the audio system which contains complementary information to MFCC. As the name suggests, MFCC is based on the mel scale whereas IMFCC is based on the inverted mel scale. Recent works showed better performance of PAD when multiple features were used as different features contain complementary information that discriminates genuine from spoof voice better [35, 48, 142]. Moreover, by using different features and models in classification through the fusion method, a significant improvement in the performance of voice PAD can be observed [15].

4.2.3 Classifiers

The third attribute of the taxonomy is classifiers. From articles surveyed, classifiers used in voice PAD can be categorized into three groups, namely conventional, deep learning, and multiple classifiers. The often-used classifiers are conventional classifiers, followed by multiple classifiers, and deep learning.

One of the widely used conventional classifiers in recent works for PAD tasks was GMM as it is an effective probabilistic model for speaker verification tasks [65]. Unlike speaker verification, UBM adaptation was not required for spoof speech detection. GMM was used to classify genuine and spoof voices in which the process is similar to that of speaker verification using GMM.

Besides, another conventional classifier known as SVM was also extensively used in recent works due to its excellent performance in classification tasks. In the recent work [66], SVM with Radial Basis Function kernel (RBF) was found to outperform classifiers such as Decision Tree, Naive Bayes, and K-Nearest Neighbour (KNN) with an EER of 1% on the evaluation set of the ASVspoof 2019 PA dataset. Several kernels for SVM were also been tested by the researchers, but none perform better than the RBF kernel. An interesting observation is that SVM with polynomial and RBF kernels produced superior detection results compared to SVM with linear kernel due to non-linearities present in 1^{st} and 2^{nd} order replay samples of the dataset.

Deep learning methods were also frequently applied in PAD tasks. Unlike conventional classifiers, deep learning is one of the machine learning that is composed of networks capable of learning without supervision from labeled data [39, 104]. From recent works, it is found that deep learning classifiers such as DNN [120], RNN [30], and CNN [88] are capable of automatic feature abstraction in which more informative features can be identified. The informative feature extracted from voices leads to better performance in voice PAD systems [88].

The application of multiple classifiers in PAD systems also can be found in the literature. In many domains involving machine learning and classification, it has been shown that applying ensemble classifiers may improve the performance of a system [42, 79]. Nevertheless, in the field of voice recognition, it has been shown that applying multiple classifiers with the same feature hardly improves the performance of the voice PAD system [15].

4.2.4 Fusion

The fourth attribute of the taxonomy is fusion. From the literature, the work may apply fusion or no fusion. With respect to those works that apply fusion, two main fusion methods were employed namely score fusion and feature fusion. Although there are other fusion methods available, the number is small.

Score fusion is undertaken such that several scores generated by voice PAD models are considered in the classification decision [91, 128]. These scores are combined using sum, max, min, mean, standard deviation, weighted or normalized sum, etc. From the literature, score fusion is the most frequently used fusion approach in the voice PAD systems and has shown effectiveness in improving the detection rate.

Feature fusion is performed either via serial feature fusion or parallel feature fusion [118] to boost the recognition rate. Serial fusion is a method of fusing features by serially combining multiple feature vector sets into a single feature vector, called a serial fused feature. Unlike serial fusion, which is based on the union-vector, parallel feature fusion is based on a complex vector, a vector that has components of complex numbers. Between these two fusion strategies, parallel feature fusion has outperformed serial feature fusion for attack detection [141].

Other fusion such as the ensemble approach was found to be more generalized against spoofing to ASV systems. A recent work [78] introduced an end-to-end ensemble approach to jointly train two models separately were perform well on LA and PA attacks. Then, a third model learned the output of the two models and yielding a single score as detection output. Experiment results showed that the ensemble approach produced EER of 9.87% and 1.75% on evaluation sets of LA and PA sets of ASVspoof 2019 dataset respectively. Though the performance on the LA task was poorer than the PA task, the ensemble result of the LA task still improvised from EER of ranged 13-16% produced by each individual model.

4.2.5 Methodology

The fifth attribute of the taxonomy is methodology. The methodology attribute can be grouped into three categories, namely classic machine learning, end-to-end learning, and hybrid approach as shown in Fig. 5. Classic machine learning is the most common method used in the surveyed work. In classic machine learning, pre-determined features that are usually manually crafted were extracted from data samples and fed into the pre-determined classifiers to predict the class label of the data samples [31]. The feature extraction and classification are two separate modules in classic machine learning. For example, the official

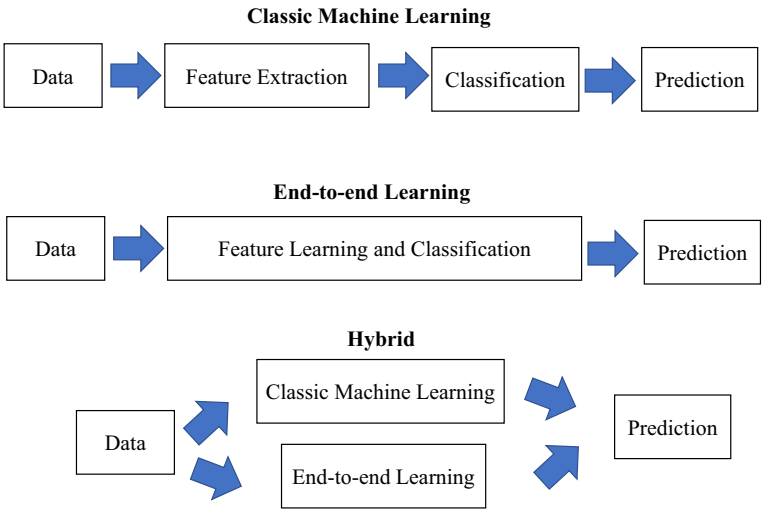


Fig. 5 Categories of the methodology used in recent voice PAD

ASVspoof baseline system is a voice PAD system based on CQCC features front-end and a GMM classifier backend [22].

In end-to-end learning, all features from data samples were identified and learned by deep learning processes automatically and jointly to determine the class label of the data samples. The feature learning and classification are under one module in end-to-end learning. Different from classic machine learning, end-to-end learning handles the entire learning process from input data to output prediction. For example, in a recent work [25], raw waveform-based deep learning spoof detection model jointly acts as both feature extractor and end-to-end classifier where there was no pre- and post-processing on the data input needed.

As the name suggests, the hybrid approach used both classic machine learning and end-to-end learning in the architecture of the voice PAD system. The hybrid methodology takes advantage of the manually crafted and automatically extracted features. The features used are varied and may provide a better representation of the data. Though the hybrid methodology is rare in the context of the PAD system, the hybrid approach has been shown to attain better performance in spoof speech detection [14].

4.2.6 Datasets

The sixth attribute of the taxonomy is datasets. The datasets attribute can be grouped into public and private datasets. There are two commonly used datasets for voice PAD researches, namely ASVspoof and AVspoof. Three Automatic Speaker Verification Spoofing and Countermeasures Challenges (ASVspoof) were organized previously, namely ASVspoof 2015, ASVspoof 2017, and ASVspoof 2019, in which datasets were made publicly available to download. ASVspoof 2015 dataset consists of speech synthesis and voice conversion attacks, whereas ASVspoof 2017 dataset consists of replay attacks. ASVspoof 2019 dataset contains speech synthesis, voice conversion, and replay attacks. As for the AVspoof dataset, it was made publicly available and it contains replay, speech synthesis, and

voice conversion attacks. Details regarding ASVspoof and AVspoof datasets can be found in [6] and [45] respectively.

ASVspoof was the most frequently used dataset among the surveyed articles. About 62% of the considered articles used ASVspoof datasets for experimentation. Some researchers applied multiple datasets for cross-database evaluation [86], but the number is limited. Furthermore, from the surveyed articles, cross-dataset experiments of voice PAD have shown that the state-of-the-art voice PAD systems were not well generalized as the performance significantly degrades when encountered unseen spoofing attacks.

4.2.7 Evaluation criteria

The last attribute of the taxonomy is the performance evaluation criteria. From the surveyed articles, it can be seen that EER is the main criterion used for performance evaluation of voice PAD systems as over three-quarters of the works evaluating their proposed PAD using EER. A limited number of the surveyed articles have chosen accuracy as the single performance evaluation criteria, for example, [44]. Some researchers evaluate their work using multiple performance evaluation criteria such as EER with min-tDCF, EER with Half Total Error Rate (HTER), and False Match Rate (FMR) with False Non-match Rate (FNMR). In 2021, the proportion of recent works that used multiple evaluation criteria was recorded at 28.57%. Since not all study was evaluated using the same criterion, this creates a problem when comparing the works. Hence, some recent works such as [62, 145], and [140] provided more than one evaluation criteria for performance comparison. A fair comparison of voice PAD in terms of performance may be made through the standardization of evaluation criteria.

4.3 An Analysis to the Trend of Voice PAD in Recent Years

This sub-section presents the analysis of voice PAD works in recent years. Statistical analyses of the trend of works on PAD is conducted to discover their limitation and subsequently project the potential future works of voice PAD. Visualizations [75] were used to show the trends based on the attributes enlisted in the taxonomy of voice PAD, described in the preceding section. Figures 6 to 20 visualize the trends; each is explained in detail.

Figures 6 and 7 visualized the analyses of the type of presentation attack attribute. Based on Fig. 6, the proportion of work on speech synthesis and voice conversion-based PAD decreases steadily from the year 2015 to 2019, whereas both replay and multiple attack type targeting PAD increases from the year 2015 to 2018. From 2019 onwards, the works on replay attack targeting PAD decreases to 14.29% in 2021 while the number of voice PAD works targeting speech synthesis and voice conversion as well as multiple types of attacks steadily increase to 42.86% and 42.86% respectively in 2021. As there is no prior knowledge regarding the types of presentation attacks for PAD systems before the detection takes place, countermeasures that can detect multiple types of attacks become the favorite. On the other hand, the decrement in the proportion of speech synthesis and voice conversion from 2015-2019 may be due to the shift in attention of researchers to replay and multiple types of attacks in the period. Nonetheless, the overall trend shift towards speech synthesis and voice conversion as well as multiple types of attacks in 2021. The increment in the proportion of multiple types of attack as shown in the statistics support the significance in detecting presentation attacks regardless of the types. Figure 6 shows that most of the past work focusing on the development of countermeasures for device-assisted attacks (98.84%). Only two works (1.16%) presented an approach to detect impersonation attacks.

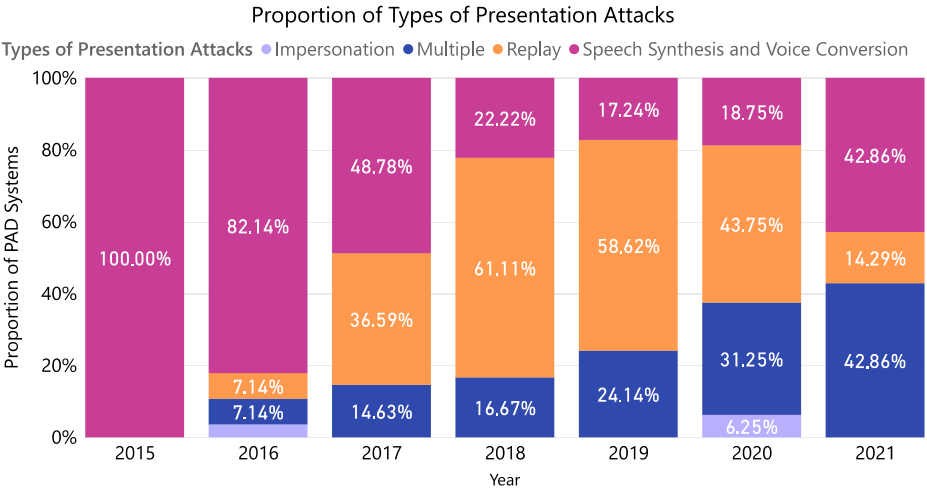


Fig. 6 The trend of the detection of types of presentation attack by voice PAD in recent years

Figures 8 and 9 visualized the analyses of the feature attribute. From the perspective of features used in recent work on voice PAD as shown in Fig. 8, the majority of works used multiple features from 2015-2021. This is because the use of multiple features contains complementary information that can be used to better discriminates genuine from spoof voice [35, 48]], [142]. However, the trend of using multiple features seems to be reduced to 57.14% in 2021. The two most frequently used individual features, CQCC and MFCC, became less preferred due to the most recent work showing the superiority of multiple features in detecting presentation attacks [20, 112, 127]. In 2021, works that applied MFCC as

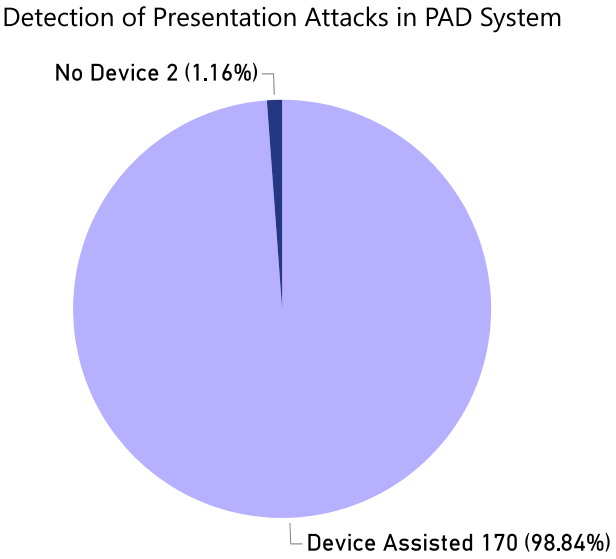


Fig. 7 The proportion of device assisted and zero-effort imposter attacks as target detection in recent voice PAD

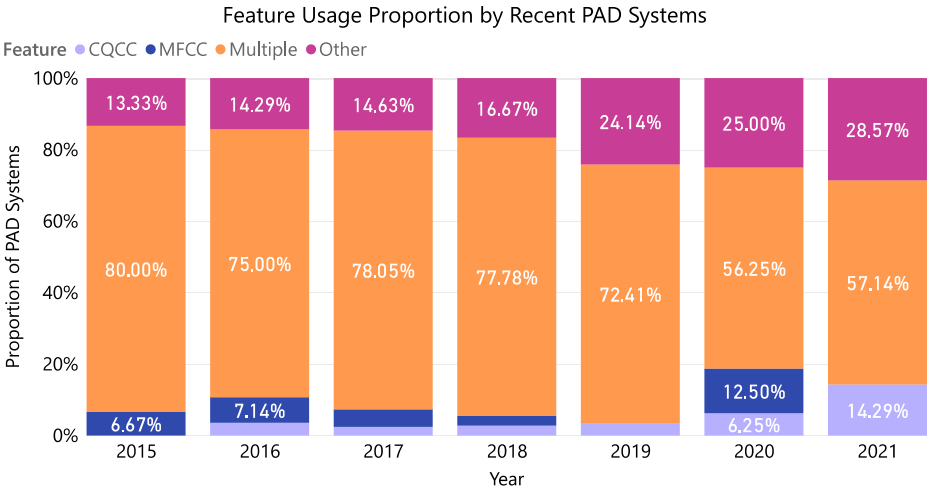


Fig. 8 The trend of features used in voice PAD in recent years

a single feature for voice PAD were none. Overall, MFCC and CQCC were used in 4.65% and 3.49% of the recent works respectively, as a single feature. Other features were used 18.02% of the total considered work.

Figures 10 and 11 visualized the analyses of the classifier attribute. As for the trend of classifiers used in voice PAD in recent years, as shown in Fig. 10, GMM was the most preferred classifier until 2018. The emergence of deep learning has impacted the selection of classifiers on voice PAD as the number of works using deep learning has steadily increased from 2016 to 2019. The utilization of multiple classifiers in the form of ensemble classifiers has also gained attention recently [37, 101]. The switched of interest on classifiers selection, concerning voice PAD, may be caused by better detection results produced by deep learning which is capable of feature abstraction while ensemble classifiers which capable of gathering complementary information over GMM [88]. Although the usage trend of GMM as an

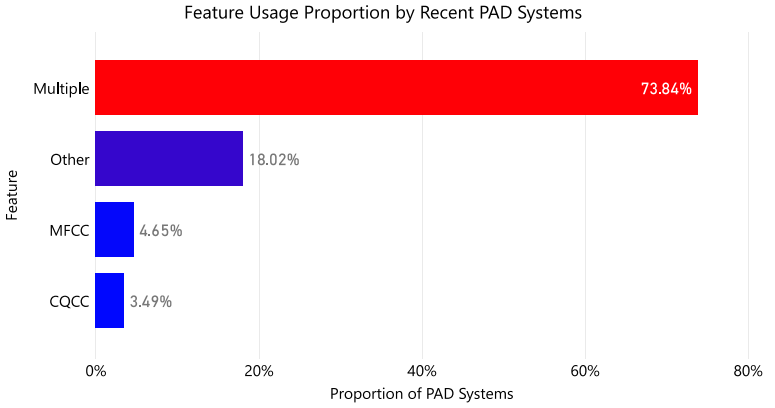


Fig. 9 The proportion of features used in recent voice PAD

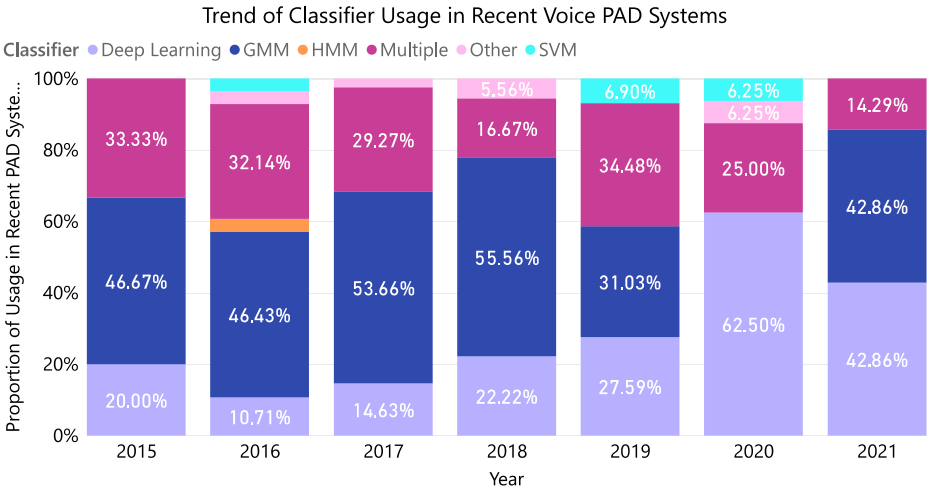


Fig. 10 The trend of classifiers used in voice PAD in recent years

individual classifier experienced fluctuation, GMM is still the most frequently used classifier (43.02%), followed by multiple classifiers (27.33%) and deep learning (23.84%) as shown in Fig. 11. The application of single classifiers like SVM, HMM, and other classifiers was very limited with shares of 2.33%, 0.58%, and 2.91% respectively in recent years.

Figures 12, 13, and 14 visualized the analyses of the fusion attribute. Figures 12 shows the trend of fusion application in a recent voice PAD. Score fusion was the most preferred fusion from 2015-201. This trend fluctuated from 2020 onwards. One interesting observation is the usage of feature fusion that receives mixed reception from works considered. Since feature fusion can produce good detection results [20, 77, 130], it is conjectured that feature fusion could still be employed in the future. Most works did not use fusion from 2020 onwards. Nonetheless, from the works considered in this paper, only 105 (61.05%) applied fusion, whereas 67 (38.95%) works did not, as shown in Fig. 13. Figure 14 shows

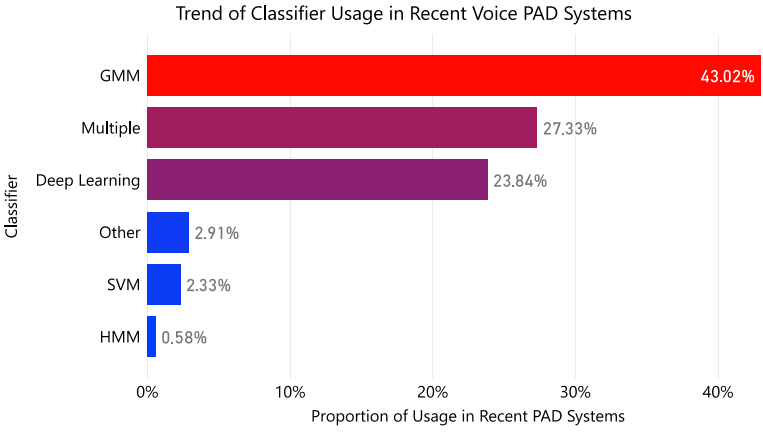


Fig. 11 the proportion of classifiers used in recent voice PAD

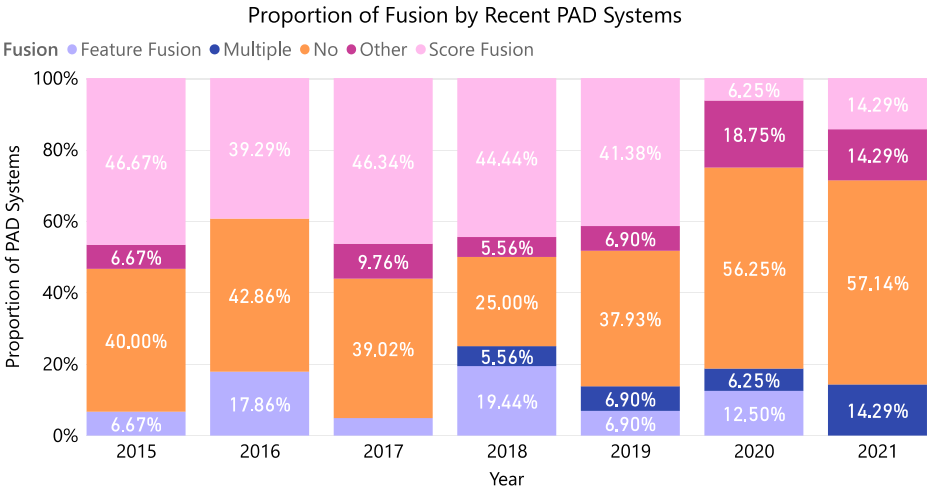


Fig. 12 The trend of the application of fusion in voice PAD in recent years

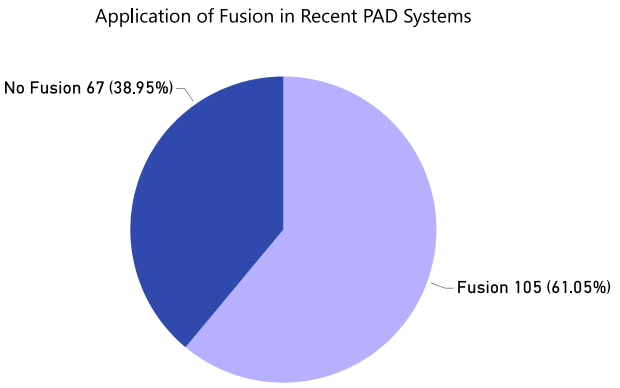


Fig. 13 The proportion of recent PAD using fusion and no fusion

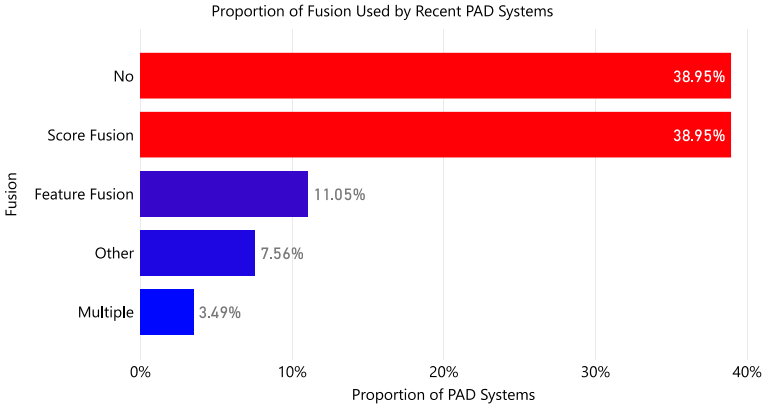


Fig. 14 The proportion of fusion used in recent voice PAD

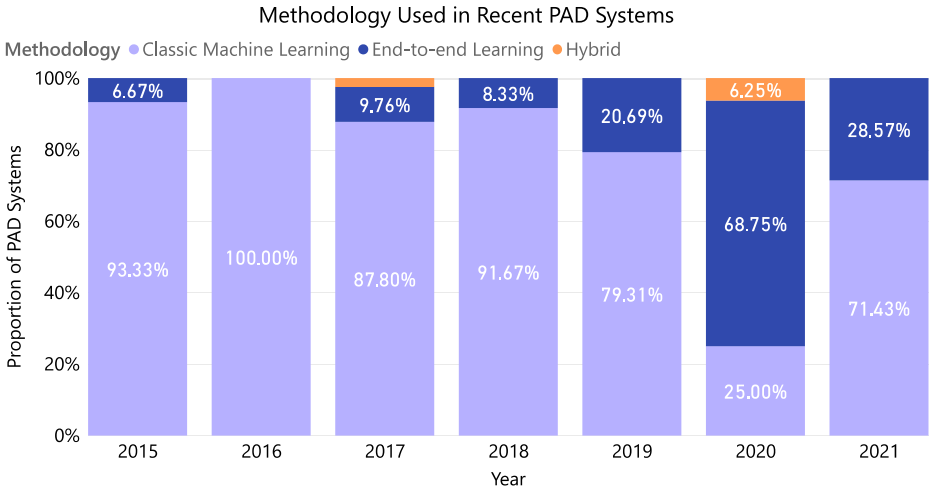


Fig. 15 The trend of the methodology used in voice PAD in recent years

that, among various fusion approaches, score fusion was used in 38.95% of the considered works. Feature fusion, multiple fusion, and other fusion methods were mere slightly exceeding 20% in usage when totaled up.

Figures 15 and 16 visualized the analyses of the methodology attribute. Figure 15 shows the trend of the methodology used in recent voice PAD systems. The classic machine learning approach was the most preferred methodology in recent years except in 2020. In 2020, end-to-end learning was the most preferable approach. There were 83.14% of recent works that contributed to voice PAD using the classic machine learning approach. This is because classic machine learning was found to outperform the end-to-end approach consistently in recent work [129]. One interesting observation is the usage of end-to-end learning that receives mixed reception from recent works such that no trend can be observed. Nonetheless, the end-to-end learning approach still contributed more than 15% in overall recent

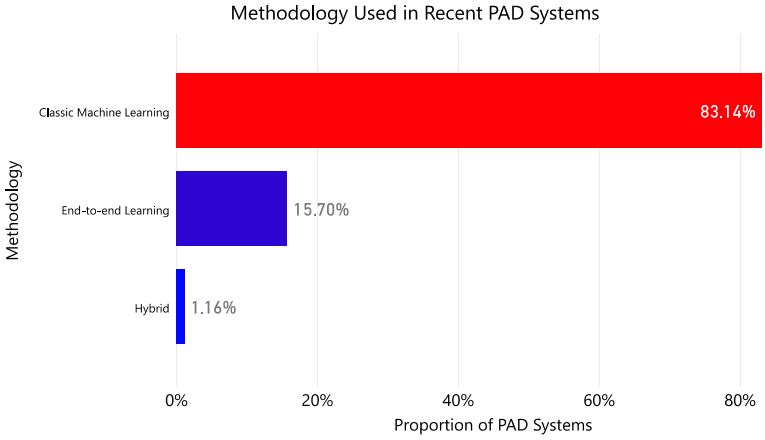


Fig. 16 The proportion of methodology used in voice PAD in recent years

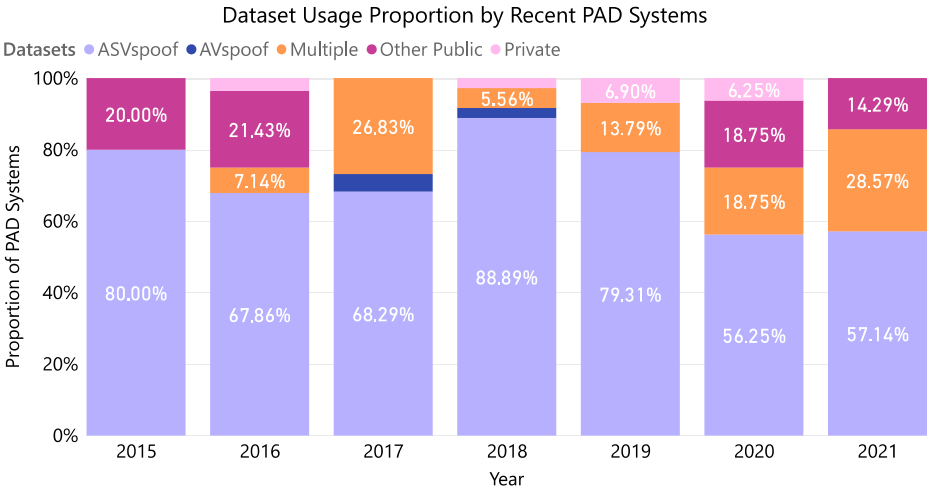


Fig. 17 The trend of datasets used in voice PAD in recent years

works considered. Only two recent works (1.16%) applied the hybrid approach in the voice PAD.

Figures 17 and 18 visualized the analyses of the datasets attribute. ASVspoof datasets were the most commonly used voice PAD datasets in recent years, recording 73.84% of usage. Multiple datasets were also used in training and evaluating recent voice PAD with a usage proportion of 13.95%. Complementarily, while only 13.95% of the recent works used multiple datasets, 86.05% of the works used a single dataset in training and evaluating the work. This trend indicates that a lack of cross-datasets evaluation was done in recent works for voice PAD, which may cause the proposed voice PAD to be less generalizable.

Figures 19 and 20 visualized the analyses of the evaluation criteria attribute. EER is the most used criteria to evaluate the performance of voice PAD with 72.67% usage across the years, as shown in Fig. 20. Though some works were using multiple evaluation criteria (23.84%), only 3.49% of the recent work evaluating the work using accuracy as the single

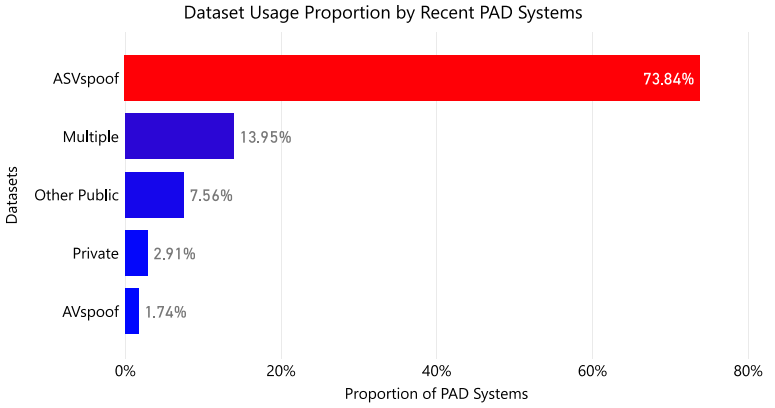


Fig. 18 The proportion of datasets used in recent voice PAD

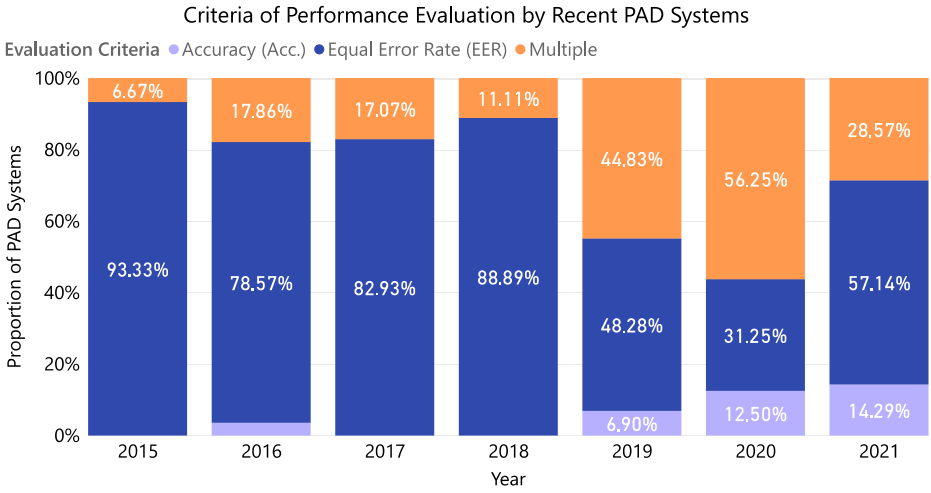


Fig. 19 The trend of evaluation criteria used in voice PAD in recent years

evaluation criteria, as shown in Fig. 20. Nonetheless, there was a dramatic decline in the use of EER as shown in Fig. 19. Meanwhile, the usage of multiple criteria was increasing steadily in recent years, except for a slight drop in 2018. Still, it experienced a significant increment to 56.25% in 2020 but dropped to 28.57% in 2021. The increase in the cumulative number of recent works that used different evaluation criteria indicates using a single metric for evaluation may not sufficient to show how well a PAD system performed.

4.4 Research Gap and Future Direction of Voice PAD

This sub-section presents the research gap and corresponding future direction of voice PAD. The state-of-the-art voice PAD field is suffering from several issues that can be found in the articles considered. Some of the proposed future works are designed to deal with the issues found. Figure 21 summarized the research gap and future direction of voice PAD. Details of the issues and potential future works are presented in Sections 4.4.1 and 4.4.2 respectively.

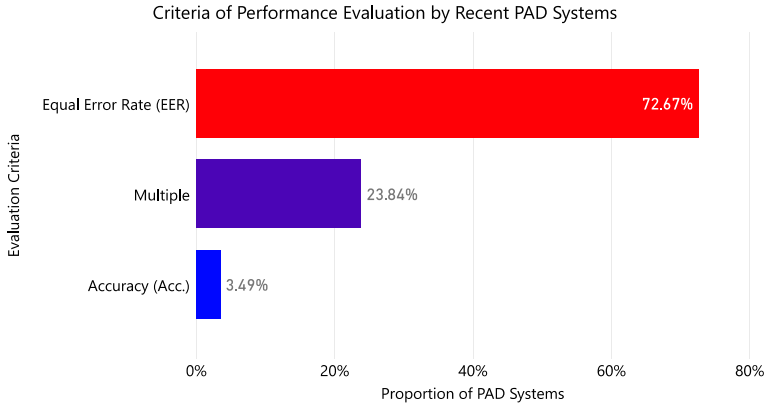


Fig. 20 The proportion of evaluation criteria used in recent voice PAD

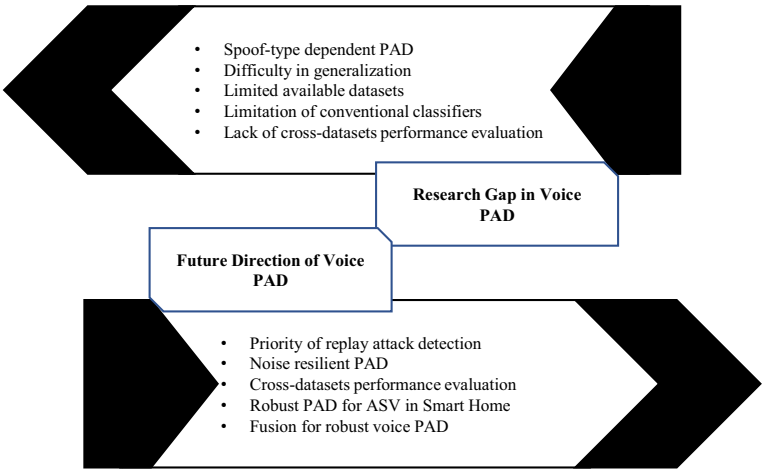


Fig. 21 Summary of research gap and future direction

4.4.1 Issues of Voice PAD

This section presents the issues of voice PAD found in the literature. Five main issues were identified: (i) spoof-type dependent PAD, (ii) difficulty in the generalization of the PAD systems, (iii) limited available datasets, (iv) limitation of conventional classifiers, and (v) lack of cross-datasets performance evaluation.

Spoof-type dependent PAD. Most of the state-of-the-art voice spoofing countermeasures are indicative and specific to the types of spoofing [28]. A PAD system that addresses speech synthesis and voice conversion may not be effective against a replay attack and vice versa. For example, a PAD system trained with genuine, speech synthesis, and voice conversion attacks dataset but tested on a dataset consisting of genuine and replays attacks only could not detect the replay attacks effectively [90].

Application of several PAD systems [80] into an ASV system may be a possible solution to compensate for the response time tradeoff. Another concern is that as most state-of-the-art PAD systems were proposed to detect specific types of spoofing attacks, security adversaries may exploit this flaw to spoof and bypass PAD systems by applying multiple spoofing types in one spoofing attempt. For example, voice conversion can be applied on top of impersonation to boost spoofing effectiveness [136]. Therefore, spoof-type dependent PAD which is designed to detect only single type of attack may not able to detect that spoofing attack.

Difficulty in generalization. State-of-the-art PAD systems are not well generalized. Several cross-database evaluations conducted by researchers have shown that the performance of voice PAD systems declined when it is trained using a dataset and tested using another dataset that has different types of spoofing attacks [32, 43, 55]. It can be seen that the current state of PAD is still dataset-dependent to achieve low error rates (FAR, FRR, HTER, and EER). The approaches in preparing different datasets by different entities are different, factors like different recording environments, recording devices, spoofing algorithms, and noise levels. As a result, the quality of different datasets of audio is different. Hence, the performance can be inconsistent when evaluating using different datasets if the PAD system is not well generalized.

Limited available datasets. The limitation of datasets availability has caused most of the PAD systems modeled using the text-independent method for voice surveillance applications [136]. Nevertheless, many ASV systems have been developed using text-dependent modeling techniques for authentication purposes [135]. In order to evaluate the performance of speaker verification with PAD, datasets that can be used for both speaker verification and PAD are needed, such that both speaker identity label and the spoof-genuine label must be made available in the datasets. The limited availability of datasets that can be used to evaluate both PAD and speaker verification has induced the limitation in validating the effectiveness of PAD systems in an actual situation.

Limitation of conventional classifiers. Conventional classifiers like GMM-UBM for speaker identification and verification are vulnerable to voice conversion attacks [84]. Since most of the current speaker verification systems are GMM based, efforts to incorporate additional steps in the GMM-based speaker verification systems to capture artificial signals are necessary [27]. This is due to conventional classifiers such as GMM-UBM and SVM do not have the capability of feature abstraction, which can be found in deeper learning classifiers such as DNN, RNN, and CNN [88]. Possible complementary information can be obtained to identify spoofing from a genuine voice better by utilizing a fusion of classifiers in different natures [88] compared to conventional classifiers.

Lack of cross-datasets performance evaluation. In a recent study, it has been shown that current voice PAD systems were not well generalized as the performance of voice PAD degrades when evaluated using different datasets [54]. To determine whether a voice PAD system is robust enough against unseen spoofing attacks, a different dataset can be used in model evaluation. The evaluation process is known as cross-datasets evaluation. However, most of the current voice PAD systems (76.51%) were evaluated on a single dataset, as shown in Fig. 18. Therefore, cross-datasets evaluations are needed to ensure the proposed voice PAD system is robust enough against unseen spoofing attacks.

4.4.2 Potential future works

This section presents the possible future works that should be considered to improve the performance of voice PAD and speaker verification.

The priority of replay attack detection. The threat level posed by the replay attack to ASV is significant. From the results of the ASVspoof 2019 Challenge, replay attacks of higher quality were difficult to be detected by state-of-the-art PAD systems [124]. In addition, most of the voice PAD systems were evaluated using a corpus made up of first-order replayed recordings (replayed once), the detection of multi-order replay attacks (replayed multiple times) has not been done [9]. Hence, more work should be directed at replay attack detection, while spoof detection should include both replay and artificial speech (speech synthesis and voice conversion). As described in Section 1, anyone can initiate a replay attack simply by using an electronic voice recorder or smartphone. No specific skills in signal processing are required to launch replay attacks. In the future, researchers may put replay attack detection as the first layer of security in ASV against presentation attacks when designing voice PAD [136].

Noise resilient PAD. Most of the performance evaluations made on proposed PAD systems consider only identical conditions for both training and testing. In reality, this is not the case. The condition of perceiving the voice input from users may vary. The variable condition to capture voice input, such as background noise, degrades the quality of voice captured by ASV systems. Similarly, in a noisy condition, the performance of PAD

systems degrade significantly [29]. Hence, future work may include a variety of background noise conditions into PAD systems performance evaluation to generate a better PAD system that is resilient to acoustic mismatch conditions [136].

Cross-datasets performance evaluation. As mentioned previously, current PAD systems are not well generalized as the performance of PAD in cross-datasets evaluation tends to degrade [54]. To verify whether the proposed PAD systems are well generalized, cross-datasets evaluation can be performed. If the voice PAD system managed to achieve consistent performance across different datasets, then the robustness of the system will be justifiable [63]. However, there were not many recent works that applied cross-datasets evaluation to show the performance of the proposed PAD systems against unseen attacks. As the technology keeps evolving, the methods to spoof ASV systems will increase. Hence, it is crucial to introduce a robust PAD system that is well generalized.

Robust PAD for ASV in Smart Home. As the concept of Smart Home promotes handsfree and automation properties [72], biometric technologies, including voice recognition, are the best method to be used for access control and personalization [125]. However, current speaker recognition systems can be vulnerable to presentation attacks as the existing PAD systems are very indicative and specific in detecting presentation attacks [28]. Hence, most of these PAD systems are inapplicable in the actual situation where types of presentation attacks are unknown in reality [28]. When applied to Smart Home, the threat of the presentation attacks would become significant. Therefore, there is an urgent need to develop a robust PAD system to secure ASV systems and hence the adoption of ASV in Smart Home applications can be accelerated.

Fusion for robust voice PAD. Last but not least, fusion can serve as an effective choice to improve the performance of voice PAD. Although there are a variety of methods to perform fusion such as score fusion and feature fusion, only 63.76% of the recent works applied fusion to enhance the performance of voice PAD. However, the increase in additional tasks for fusion computation leads to the increase of computation time of voice PAD, which may be served as the reason for 36.24% of the recent works not apply fusion. Nonetheless, the trade-off between computation time and the detection rate should be reviewed in the future whether the enhancement in the detection rate worth the trade-off computation time induced by the fusion.

5 Conclusion

As time progress, more works on voice PAD were published. However, there was a limited systematic survey available on the current state of research and application. To the best of our knowledge, most of the papers did not provide a detailed taxonomy of recent voice PADs. This paper is thus produced to offer an extensive survey of speaker verification systems, spoofing attacks, and voice PAD to secure speaker verification systems published in 2015 to 2019. A total of 172 Scopus indexed articles on voice PAD were considered in producing this survey.

In order to understand the trend of work on voice PAD systems, a taxonomy of state-of-the-art voice PAD systems was built based on the survey on the works considered. Analyses of the trend on recent works on PAD, based on the identified attributes from the taxonomy, were also presented. From the analyses, the researchers' interest in developing models to detect multiple types of attacks is increasing. Furthermore, deep learning usage as classifiers for voice PAD has also increased since 2016, although GMM is still the most frequently used classifier.

The research gap and future direction of voice PAD were subsequently established and described in this paper. There were five existing issues of voice PAD identified, namely spoof-type dependent PAD, difficulty in generalization, limited available datasets, limitation of conventional classifiers, and lack of cross-datasets performance evaluation. Five potential works for voice PAD were suggested to resolve the identified issues, namely priority of replay attack detection, noise resilient PAD, cross-datasets performance evaluation, robust PAD for ASV in Smart Home, and fusion for robust voice PAD.

To conclude, investigating how voice PAD was employed in ASV systems is highly significant to ensure future research will concentrate on the right dimension of the voice PAD, thereby improving voice PAD systems performance. The presented taxonomy could be used by other researchers to plan their research contributions and activities. The potential future direction found could further enhance efficiency and increase the number of voice PAD systems applications.

Author Contributions All authors contributed to the study conception and design. Material preparation and analysis were performed by Choon Beng Tan and Mohd Hanafi Ahmad Hijazi. The first draft of the manuscript was written by Choon Beng Tan, and supervised by Mohd Hanafi Ahmad Hijazi. All authors provided critical feedback and helped shape the analysis and manuscript.

Availability of data and material The authors declare that data and material used to write this paper can be found with provided references.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abozaid A, Haggag A, Kasban H, Eltokhy M (2018) Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-018-7012-3
2. Adel M, Afify M, Gaballah A (2018) Text-Independent Speaker Verification Based on Deep Neural Networks and Segmental Dynamic Time Warping. 2018 IEEE Spoken Language Technology Workshop (SLT), pp 1001–1006, 1806.09932
3. Adiban M, Sameti H, Shehnepoor S (2020) Replay spoofing countermeasure using autoencoder and siamese networks on ASVspoof 2019 challenge. Computer Speech & Language 64:101105. https://doi.org/10.1016/j.csl.2020.101105
4. Admuthe SS, Ghugardare S (2015) Survey paper on automatic speaker recognition systems. In: International conference on multimedia, computer graphics, and broadcasting international conference on signal processing, image processing, and pattern recognition, vol 4, pp 10895-10898
5. Al-Ali AKH, Senadji B, Naik GR (2017) Enhanced forensic speaker verification using multi-run ICA in the presence of environmental noise and reverberation conditions. In: 2017 IEEE International conference on signal and image processing applications (ICSIPA), IEEE, pp 174–179. https://doi.org/10.1109/ICSIPA.2017.8120601

6. ASVspoof (2019) ASVspoof 2019 Automatic Speaker Verification Spoofing and Countermeasures Challenge. https://www.asvspoof.org/
7. ASVspoof consortium (2019) ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019, pp 1–19
8. Auckenthaler R, Carey M, Lloyd-Thomas H (2000) Score normalization for text-independent speaker verification systems. Digital Signal Processing: A Review Journal 10(1):42–54. https://doi.org/10.1006/dspr.1999.0360
9. Baumann R, Malik KM, Javed A, Ball A, Kujawa B, Malik H (2021) Voice spoofing detection corpus for single and multi-order audio replays. Computer Speech & Language 65:101132. https://doi.org/10.1016/j.csl.2020.101132
10. Billal K, Abdelhakim D (2017) A new speaker verification algorithm based on identification results. In: 2017 5Th international conference on electrical engineering - boumerdes (ICEE-B), IEEE, pp 1–6. https://doi.org/10.1109/ICEE-B.2017.8192139
11. Biometrics TF (2008) Biometrics Glossary (BG). https://www.hsdl.org/?view&did=32101
12. Biometrics Institute (2017) Types of Biometrics. https://www.biometricsinstitute.org/types-of-biometrics
13. Bonifaco H, Guzman KR, Jara JN, Jasareno AD, Zabala AC, Prado SV, Buenaventura CS (2017) Comparative analysis of filipino-based rhinolalia aperta speech using mel frequency cepstral analysis and Perceptual Linear Prediction. In: 2017IEEE 9Th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM), IEEE, pp 1–6. https://doi.org/10.1109/HNICEM.2017.8269507
14. Cai W, Cai D, Liu W, Li G, Li M (2017) Countermeasures for automatic speaker verification replay spoofing attack : on data augmentation, feature representation, classification and fusion. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 17–21. https://doi.org/10.21437/Interspeech.2017-906
15. Chen Z, Xie Z, Zhang W, Xu X (2017) Resnet and Model Fusion for Automatic Spoofing Detection. In: Interspeech 2017, ISCA, ISCA, pp 102–106. https://doi.org/10.21437/Interspeech.2017-1085
16. Chen Z, Zhang W, Xie Z, Xu X, Chen D (2018) Recurrent neural networks for automatic replay spoofing attack detection. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 2052–2056. https://doi.org/10.1109/ICASSP.2018.8462644
17. Chettri B, Sturm BL (2018) A deeper look at gaussian mixture model based Anti-Spoofing systems. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, IEEE, pp 5159–5163. https://doi.org/10.1109/ICASSP.2018.8461467
18. Chettri B, Mishra S, Sturm BL, Benetos E (2018) Analysing The Predictions Of a CNN-based Replay Spoofing Detection System. In: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp 92–97. https://doi.org/10.1109/SLT.2018.8639666
19. Chettri B, Benetos E, Sturm BLT (2020) Dataset Artefacts in Anti-Spoofing systems: A Case Study on the ASVspoof 2017 Benchmark. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28:3018–3028. https://doi.org/10.1109/TASLP.2020.3036777
20. Das RK, Yang J, Li H (2019) Long range acoustic features for spoofed speech detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 1058–1062. https://doi.org/10.21437/Interspeech.2019-1887
21. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2011) Front-End Factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing 19(4):788–798
22. Delgado H, Todisco M, Sahidullah M, Evans N, Kinnunen T, Lee KA, Yamagishi J (2018) ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In: Odyssey 2018 - The Speaker and Language Recognition Workshop, pp 296–303. https://doi.org/10.21437/odyssey.2018-42
23. Demiroglu C, Buyuk O, Khodabakhsh A, Maia R (2017) Postprocessing synthetic speech with a complex cepstrum vocoder for spoofing phase-based synthetic speech detectors. IEEE J Select Top Signal Process 11(4):671–683. https://doi.org/10.1109/JSTSP.2017.2673807
24. Dey S, Koshinaka T, Motlicek P, Madikeri S (2018) DNN based speaker embedding using content information for Text-Dependent speaker verification. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5344–5348. https://doi.org/10.1109/ICASSP.2018.8461389
25. Dinkel H, Chen N, Qian Y, Yu K (2017) End-to-end spoofing detection with raw waveform CLDNNS. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4860–4864. https://doi.org/10.1109/ICASSP.2017.7953080
26. Dua M, Jain C, Kumar S (2021) LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-021-02960-0

27. Evans N, Alegre F, Wu Z, Kinnunen T (2009) Encyclopedia of biometrics. Springer, Boston. https://doi.org/10.1007/978-3-642-27733-7
28. Evans N, Kinnunen T, Yamagishi J, Wu Z, Alegre F, Leon PD (2014) Speaker Recognition Anti- Spoofing. Handbook of Biometric Anti-Spoofing pp 125–146. https://doi.org/10.1007/978-1-4471-6524-8
29. Gomez-alanis A, Peinado AM, Gonzalez JA, Gomez AM (2018) A Deep Identity Representation for Noise Robust Spoofing Detection. In: Interspeech 2018, September, pp 676–680
30. Gomez-Alanis A, Gonzalez-Lopez JA, Peinado AM (2020) A Kernel Density Estimation Based Loss Function and its Application to ASV-Spoofing Detection. IEEE Access 8:108530–108543. https://doi.org/10.1109/ACCESS.2020.3000641
31. Gomez-Alanis A, Gonzalez-Lopez JA, Dubagunta SP, Peinado AM, Magimai-Doss M (2021) On joint optimization of automatic speaker verification and Anti-Spoofing in the embedding space. IEEE Trans Inform Forensics Secur 16:1579–1593. https://doi.org/10.1109/TIFS.2020.3039045
32. Goncalves AR, Violato RP, Korshunov P, Marcel S, Simoes FO (2017) On the generalization of fused systems in voice presentation attack detection. In: 2017 International conference of the biometrics special interest group (BIOSIG), IEEE, pp 1–5. https://doi.org/10.23919/BIOSIG.2017.8053516
33. Gong Y, Yang J, Huber J, MacKnight M, Poellabauer C (2019) REMASC: Realistic replay attack corpus for voice controlled systems. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2355–2359. https://doi.org/10.21437/Interspeech.2019-1541, arXiv:1904.03365v2
34. Gong Y, Yang J, Poellabauer C (2020) Detecting replay attacks using multi-channel audio: A neural network-based method. IEEE Signal Process Lett 27:920–924. https://doi.org/10.1109/LSP.2020.2996908, 2003.08225
35. Hanilci C (2017) Speaker verification anti-spoofing using linear prediction residual phase features. In: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, pp 96–100. https://doi.org/10.23919/EUSIPCO.2017.8081176
36. Hanilçi C (2018) Data selection for i-vector based automatic speaker verification anti-spoofing. Digital Signal Process Rev J 72:171–180. https://doi.org/10.1016/j.dsp.2017.10.010
37. Hanilçi C (2018) Features and classifiers for replay spoofing attack detection. In: 2017 10Th international conference on electrical and electronics engineering, ELECO 2017, pp 1187–1191
38. Hanilçi C, Kinnunen T, Sahidullah M, Sizov A (2015) Classifiers for synthetic speech detection: A comparison. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2057–2061
39. Haviluddin H, Alfred R, Obit J, Hijazi MHA, Ibrahim AAA (2015) A performance comparison of statistical and machine learning techniques in learning time series data. Adv Sci Lett 21(10):3037–3041. https://doi.org/10.1166/asl.2015.6490
40. Heigold G, Moreno I, Bengio S, Shazeer N (2018) End-to-End text-dependent speaker verification. In: Acoustics, speech, and signal processing (ICASSP), International Conference, pp 3–7
41. Hemavathi R, Kumaraswamy R (2021) Voice conversion spoofing detection by exploring artifacts estimates. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-10212-0
42. Hijazi MHA, Beng TC, Mountstephens J, Yuto L, Nisar K (2018) Malware Classification Using Ensemble Classifiers. Advanced Sci Lett 24(2):1172–1176. https://doi.org/10.1166/asl.2018.10710
43. Himawan I, Villavicencio F, Sridharan S, Fookes C (2019) Deep domain adaptation for anti-spoofing in speaker verification systems. Computer Speech and Language 58:377–402. https://doi.org/10.1016/j.csl.2019.05.007
44. Huang T, Wang H, Chen Y, He P (2020) GRU-SVM Model for Synthetic Speech Detection. In: Digital Forensics and Watermarking, pp 115–125. https://doi.org/10.1007/978-3-030-43575-2
45. Idiap Dataset Distribution Portal (2015) The AVspoof Database. https://www.idiap.ch/dataset/avspoof
46. Jiang X, Wang S, Xiang X, Qian Y (2018) Integrating online i-vector into GMM-UBM for text-dependent speaker verification. In: Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, pp 1628–1632. https://doi.org/10.1109/APSIPA.2017.8282293
47. Jin M, Yoo CD (2010) Speaker verification and identification. Behavioral Biometrics for Human Identification, pp 264–289. https://doi.org/10.4018/978-1-60566-725-6.ch013
48. Kamble MR, Patil HA (2018) Novel energy separation based frequency modulation features for spoofed speech classification. In: 2017 9th International Conference on Advances in Pattern Recognition, ICAPR 2017, IEEE, pp 326–331. https://doi.org/10.1109/ICAPR.2017.8593041
49. Kamble MR, Sailor HB, Patil HA, Li H (2019) Advances in anti-spoofing: From the perspective of ASVspoof challenges. APSIPA Transactions on Signal and Information Processing 9. https://doi.org/10.1017/ATSIP.2019.21

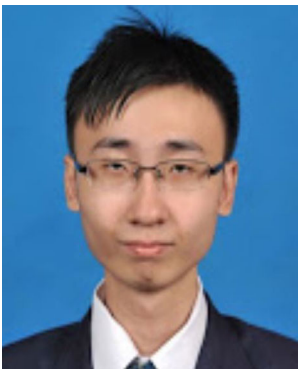
50. Kinnunen T, Evans N, Yamagishi J, Lee KA, Todisco M (2017) ASVSpoof 2017 : Automatic speaker verification spoofing and countermeasures challenge evaluation plan. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2017, pp 1–6
51. Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee KA (2017) The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2017, pp 2–6. https://doi.org/10.21437/Interspeech.2017-1111
52. Kinnunen T, Sahidullah M, Falcone M, Costantini L, Hautamäki RG, Thomsen D, Sarkar A, Tan ZH, Delgado H, Todisco M, Evans N, Hautamäki V, Lee KA (2017) RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In: ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp 5395–5399. https://doi.org/10.1109/ICASSP.2017.7953187
53. Kinnunen T, Lee KA, Delgado H, Evans N, Todisco M, Sahidullah M, Yamagishi J, Reynolds DA (2018) t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. In: Odyssey 2018 The Speaker and Language Recognition Workshop, pp 312–319. https://doi.org/10.21437/odyssey.2018-44, 1804.09618
54. Korshunov P, Marcel S (2016) Cross-database evaluation of audio-based spoofing detection systems. In: INTERSPEECH 2016, pp 1705–1709
55. Korshunov P, Marcel S (2017) Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations. IEEE J Select Top Signal Process 11(4):695–705. https://doi.org/10.1109/JSTSP.2017.2692389
56. Korshunov P, Marcel S (2017) Presentation attack detection in voice biometrics. In: Vielhauer C (ed)
57. Kotta H, Patil AT, Acharya R, Patil HA (2020) Subband channel selection using teo for replay spoof detection in voice assistants. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp 538–542
58. Kumar AK, Paul D, Pal M, Sahidullah M, Saha G (2021) Speech frame selection for spoofing detection with an application to partially spoofed audio-data. Int J Speech Technol 24(1):193–203. https://doi.org/10.1007/s10772-020-09785-w
59. Lai CI, Chen N, Villalba J, Dehak N (2019) ASSERT: Anti-spoofing with squeeze-excitation and residual networks. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 1013–1017. https://doi.org/10.21437/Interspeech.2019-1794, 1904.01120
60. Lavrentyeva G, Novoselov S, Malykh E, Kozlov A, Kudashev O, Shchemelinin V (2017) Audio Replay Attack Detection with Deep Learning Frameworks. In: Interspeech 2017, ISCA, ISCA, vol 2017-Augus, pp 82–86. https://doi.org/10.21437/Interspeech.2017-360
61. Lee KA, Larcher A, Wang G, Kenny P, Brümmer N, Van Leeuwen D, Aronowitz H, Kockmann M, Vaquero C, Ma B, Li H, Stafylakis T, Alam J, Swart A, Perez J (2015) The RedDots data collection for speaker recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2996–3000
62. Lei Z, Yang Y, Liu C, Ye J (2020) Siamese convolutional neural network using gaussian probability feature for spoofing speech detection. In : Interspeech 2020, ISCA, ISCA, pp 1116–1120. https://doi.org/10.21437/Interspeech.2020-2723
63. Li J, Sun M, Zhang X, Wang Y (2020) Joint decision of Anti-Spoofing and automatic speaker verification by Multi-Task learning with contrastive loss. IEEE Access 8:7907–7915. https://doi.org/10.1109/ACCESS.2020.2964048
64. Li L, Chen Y, Shi Y, Tang Z, Wang D (2017) Deep speaker feature learning for text-independent speaker verification. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 1542–1546. https://doi.org/10.21437/Interspeech.2017-452, 1705.03670
65. Li SZ, Zhang D, Ma C, Shum HY, Chang E (2003) Learning to boost GMM based speaker verification. In: EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology, pp 1677–1680
66. Malik KM, Javed A, Malik H, Irtaza A (2020) A light-weight replay detection framework for voice controlled IoT devices. IEEE J Select Top Signal Process 14(5):982–996. https://doi.org/10.1109/JSTSP.2020.2999828
67. Mallouh AA, Qawaqneh Z, Barkana BD (2018) New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification. Neural Comput Applic 30(8):2581–2593. https://doi.org/10.1007/s00521-017-2848-4
68. Mariethoz J, Bengio S (2006) Can a Professional Imitator Fool a GMM-Based Speaker Verification System? Tech. rep. LIDIAP

69. Markowitz J, Markowitz J, Road NS (2008) Speaker identification and verification (SIV) applications and markets. Tech. rep., VoiceXML
70. Matějka P, Novotný O, Plchot O, Burget L, Sánchez MD, Černocký JH Analysis of score normalization in multilingual speaker recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 1567–1571. https://doi.org/10.21437/Interspeech.2017-803
71. Mather F (2017) From Scotland Yard to touchless authentication – fingerprinting makes its mark. Biometric Technology Today 2017(3):7–9. https://doi.org/10.1016/S0969-4765(17)30055-3
72. Matic M, Stefanovic I, Radosavac U, Vidakovic M (2017) Challenges of integrating smart home automation with cloud based voice recognition systems. In: 2017 IEEE 7Th international conference on consumer electronics - berlin (ICCE-Berlin), IEEE, pp 248–249. https://doi.org/10.1109/ICCE-Berlin.2017.8210640
73. Mayhew S (2015) History of Biometrics. https://www.biometricupdate.com/201802/history-of-biometrics-2
74. McGettigan C, Eisner F, Agnew ZK, Manly T, Wisbey D, Scott SK (2013) T'ain't What You Say, It's the Way That You Say It —Left Insula and Inferior Frontal Cortex Work in Interaction with Superior Temporal Regions to Control the Performance of Vocal Impersonations. Journal of Cognitive Neuroscience 25(11):1875–1886. 1511.04103
75. Mehta N, Pandit A, Shukla S (2019) Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study. J Biomed Inform 100(November 2018):103311. https://doi.org/10.1016/j.jbi.2019.103311
76. Mekonnen BW, Derebssa Dufera B (2015) Noise robust speaker verification using GMM-UBM multi-condition training. In: IEEE AFRICON Conference, IEEE, pp 1–5. https://doi.org/10.1109/AFRCON.2015.7331916
77. Mishra J, Singh M, Pati D (2018) Processing linear prediction residual signal to counter replay attacks. In: 2018 International conference on signal processing and communications (SPCOM), IEEE, pp 95–99. https://doi.org/10.1109/SPCOM.2018.8724390
78. Monteiro J, Alam J, Falk TH (2020) An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers. In: ICASSP 2020 - 2020 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6599–6603. https://doi.org/10.1109/ICASSP40776.2020.9054558
79. Monteiro J, Alam J, Falk TH (2020) Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. Computer Speech & Language 63:101096. https://doi.org/10.1016/j.csl.2020.101096
80. Muckenhirn H, Magimai-Doss M, Marcel S (2018) End-to-End convolutional neural network-based voice presentation attack detection. In: IEEE International Joint Conference on Biometrics, IJCB 2017, vol 2018-Janua, pp 335–341. https://doi.org/10.1109/BTAS.2017.8272715
81. Muhammad G, Alhamid MF, Alsulaiman M, Gupta B (2018) Edge computing with cloud for voice disorder assessment and treatment. IEEE Commun Mag 56(4):60–65. https://doi.org/10.1109/MCOM.2018.1700790
82. Nagarsheth P, Khoury E, Patil K, Garland M (2017) Replay attack detection using DNN for channel discrimination. In: Interspeech 2017, ISCA, ISCA, pp 97–101. https://doi.org/10.21437/Interspeech.2017-1377
83. Neelima M, Santiprabha I (2020) Mimicry voice detection using convolutional neural networks. In: 2020 International conference on smart electronics and communication (ICOSEC), IEEE, pp 314–318. https://doi.org/10.1109/ICOSEC49089.2020.9215407
84. Pal M, Saha G (2015) On robustness of speech based biometric systems against voice conversion attack. Appl Soft Comput J 30:214–228. https://doi.org/10.1016/j.asoc.2015.01.036
85. Pal M, Paul D, Saha G (2018) Synthetic speech detection using fundamental frequency variation and spectral features. Computer Speech and Language 48:31–50. https://doi.org/10.1016/j.csl.2017.10.001
86. Parasu P, Epps J, Sriskandaraja K, Suthokumar G (2020) Investigating Light-ResNet architecture for spoofing detection under mismatched conditions. In: Interspeech 2020, ISCA, ISCA, pp 1111–1115. https://doi.org/10.21437/Interspeech.2020-2039
87. Patel TB, Patil HA (2015) Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech . In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, ISCA, pp 2062–2066
88. Patil HA, Kamble MR (2018) A survey on replay attack detection for automatic speaker verification (ASV) system. In: 2018 Asia-pacific signal and information processing association annual summit and conference, APSIPA ASC, IEEE, pp 1047–1053. https://doi.org/10.23919/APSIPA.2018.8659666

89. Paul D, Pal M, Saha G (2017) Spectral features for synthetic speech detection. IEEE J Select Top Signal Process 11(4):605–617. https://doi.org/10.1109/JSTSP.2017.2684705
90. Paull D, Saha G (2017) Generalization of Spoofing Countermeasures: A Case Study with ASVspoof 2015 And BTAS 2016 Corpora. In: ICASSP2017, pp 2047–2051
91. Peng X, Wang L, Wang X, Qiao Y (2015) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Comput Vis Image Underst 150:109–125. https://doi.org/10.1016/j.cviu.2016.03.013
92. Prajapati GP, Kamble MR, Patil HA (2021) Energy separation based features for replay spoof detection for voice assistant. In: 2020 28Th european signal processing conference (EUSIPCO), IEEE, pp 386–390. https://doi.org/10.23919/Eusipco47968.2020.9287577
93. Rahmeni R, Aicha AB, Ayed YB (2020) Speech spoofing detection using SVM and ELM technique with acoustic features. In: 2020 5Th international conference on advanced technologies for signal and image processing (ATSIP), IEEE, pp 1–4. https://doi.org/10.1109/ATSIP49331.2020.9231799
94. Ramgire JB, Jagdale PSM (2016) A survey on speaker recognition with various feature extraction and classification techniques. Int Res J Eng Technol (IRJET) 3(4):709–712
95. Reynolds DA (1995) Speaker identification and verification using Gaussian mixture speaker models. Speech Comm 17(1-2):91–108. https://doi.org/10.1016/0167-6393(95)00009-D
96. Reynolds DA (2009) Gaussian mixture models. In: Encyclopedia of biometrics. Springer, Boston, pp 659–663
97. Reynolds DA, Rose R (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 3(1):72–83. https://doi.org/10.1109/89.365379
98. Ross A, Jain AK, Nandakumar K (2006) Score level fusion. In: Handbook of multibiometrics. Kluwer Academic Publishers, Boston, pp 91–142
99. Rupesh Kumar S, Bharathi B (2021) A novel approach towards generalization of countermeasure for spoofing attack on ASV systems. Circuits, Systems, and Signal Processing 40(2):872–889. https://doi.org/10.1007/s00034-020-01501-y
100. Sabhanayagam T, Prasanna Venkatesan V, Senthamaraikannan K (2018) A comprehensive survey on various biometric systems. Int J Appl Eng Res 13(5):2276–2297
101. Safavi S, Gan H, Mporas I (2017) Improving speaker verification performance under spoofing attacks by fusion of different operational modes. In: Proceedings - 2017 IEEE 13th International Colloquium on Signal Processing and its Applications, CSPA 2017, pp 219–223. https://doi.org/10.1109/CSPA.2017.8064954
102. Sahidullah M, Kinnunen T, Hanilçi C (2015) A comparison of features for synthetic speech detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2087–2091
103. Sahidullah M, Delgado H, Todisco M, Kinnunen T, Evans N, Yamagishi J, Lee KA (2019) Introduction to voice presentation attack detection and recent advances. Springer International Publishing, pp 321–361
104. Sailor HB, Kamble MR, Patil HA (2017) Unsupervised representation learning using convolutional restricted boltzmann machine for spoof speech detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2601–2605. https://doi.org/10.21437/Interspeech.2017-1393
105. Sanchez J, Saratxaga I, Hernaez I, Navas E, Erro D, Raitio T (2015) Toward a universal synthetic speech spoofing detection using phase information. IEEE Transactions on Information Forensics and Security 10(4):810–820. https://doi.org/10.1109/TIFS.2015.2398812
106. Saratxaga I, Sanchez J, Wu Z, Hernaez I, Navas E (2016) Synthetic speech detection using phase information. Speech Comm 81:31–41. https://doi.org/10.1016/j.specom.2016.04.001
107. Sarkar AK, Tan ZH (2016) Text dependent speaker verification using un-supervised HMM-UBM and Temporal GMM-UBM. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol 08-12-Sept, pp 425–429. https://doi.org/10.21437/Interspeech.2016-362
108. Sarria-Paja M, Senoussaoui M, O'Shaughnessy D, Falk TH (2016) Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification. In: 2016 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5480–5484. https://doi.org/10.1109/ICASSP.2016.7472725
109. Sharma V, Bansal PK (2013) A review on speaker recognition approaches and challenges. Int J Eng Res Technol 2(5):1581–1588
110. Hj Shim, Heo HS, Jw Jung, Yu HJ (2020) Self-Supervised Pre-Training With acoustic configurations for replay spoofing detection. In: Interspeech 2020, ISCA, ISCA, pp 1091–1095. https://doi.org/10.21437/Interspeech.2020-1345

111. Simmons D (2017) BBC fools HSBC voice recognition security system. https://www.bbc.com/news/technology-39965545
112. Singh M, Pati D (2019) Combining evidences from Hilbert envelope and residual phase for detecting replay attacks. Int J Speech Technol 22(2):313–326. https://doi.org/10.1007/s10772-019-09604-x
113. Singh R, Jiménez A (2017) Voice disguise by mimicry: deriving statistical articulometric evidence to evaluate claimed impersonation. IET Biometrics 6(4):282–289. https://doi.org/10.1049/iet-bmt.2016.0126
114. Sinitca AM, Efimchik NV, Shalugin ED, Toropov VA, Simonchik K (2020) Voice antispoofing system vulnerabilities research. In: 2020 IEEE Conference of russian young researchers in electrical and electronic engineering (EIConRus), IEEE, pp 505–508. https://doi.org/10.1109/EIConRus49466.2020.9039393
115. Snyder D, Garcia-Romero D, Povey D, Khudanpur S (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. In: Interspeech 2017, ISCA, ISCA, vol 2017-Augus, pp 999–1003. https://doi.org/10.21437/Interspeech.2017-620
116. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) x-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375
117. Sujiya S, Chandra E (2017) A review on speaker recognition. Int J Eng Technol 9(3):1592–1598. https://doi.org/10.21817/ijet/2017/v9i3/170903513
118. Sun QS, Zeng SG, Liu Y, Heng PA, Xia DS (2005) A new method of feature fusion and its application in image recognition. Pattern Recogn 38(12):2437–2448. https://doi.org/10.1016/j.patcog.2004.12.013
119. Suthokumar G, Sriskandaraja K, Sethu V, Wijenayake C, Ambikairajah E (2018) An Investigation about the Scalability of the Spoofing Detection System. In: 2018 IEEE 9th International Conference on Information and Automation for Sustainability, ICIAfS 2018, IEEE, pp 1–5. https://doi.org/10.1109/ICIAFS.2018.8913369
120. Suthokumar G, Sethu V, Sriskandaraja K, Ambikairajah E (2020) Adversarial Multi-Task learning for speaker normalization in replay detection. In: ICASSP 2020 - 2020 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6609–6613. https://doi.org/10.1109/ICASSP40776.2020.9054322
121. Tieran Z, Jiqing H, Guibin Z (2018) Deep neural network based discriminative training for i-vector/PLDA speaker verification. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5354–5358. https://doi.org/10.1109/ICASSP.2018.8461344
122. Todisco M, Delgado H, Evans N (2016) A new feature for automatic speaker verification anti-spoofing: Constant Q Cepstral coefficients. In: Odyssey 2016, pp 283–290. https://doi.org/10.21437/odyssey.2016-41
123. Todisco M, Delgado H, Evans N (2017) Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech and Language 45(September 2017):516–535. https://doi.org/10.1016/j.csl.2017.01.001
124. Todisco M, Wang X, Vestman V, Nautsch A, Yamagishi J, Evans N, Kinnunen T, Lee KA (2019) ASVspoof 2019 : Future horizons in spoofed and fake audio detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019, pp 3–7. arXiv:1904.05441v2
125. Tsai WH, Lin JC, Ma CH, Liao YF (2016) Speaker identification for personalized smart TVs. In: 2016 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016, IEEE, pp 1–2. https://doi.org/10.1109/ICCE-TW.2016.7521051
126. Valin JM, Skoglund J (2019) LPCNET: improving neural speech synthesis through linear prediction. In: ICASSP 2019 - 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5891–5895. https://doi.org/10.1109/ICASSP.2019.8682804
127. Villalba J, Miguel A, Ortega A, Lleida E (2015) Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2067–2071
128. Vishi K, Mavroeidis V (2018) An evaluation of score level fusion approaches for fingerprint and finger-vein biometrics. arXiv:abs/1805.1:1–11, 1805.10666
129. Wang D, Li L, Tang Z, Zheng TF (2018) Deep speaker verification: Do we need end to end? In: Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, pp 177–181. https://doi.org/10.1109/APSIPA.2017.8282024, arXiv:1706.07859v1
130. Wang L, Yoshida Y, Kawakami Y, Nakagawa S (2015) Relative phase information for detecting human speech and spoofed speech. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 2092–2096

131. Wang X, Yamagishi J, Todisco M, Delgado H, Nautsch A, Evans N, Sahidullah M, Vestman V, Kinnunen T, Lee KA, Juvela L, Alku P, Peng YH, Hwang HT, Tsao Y, Wang HM, Maguer SL, Becker M, Henderson F, Clark R, Zhang Y, Wang Q, Jia Y, Onuma K, Mushika K, Kaneda T, Jiang Y, Liu LJ, Wu YC, Huang WC, Toda T, Tanaka K, Kameoka H, Steiner I, Matrouf D, Bonastre JF, Govender A, Ronanki S, Zhang JX, Ling ZH (2019) ASVspoof 2019: a large-scale public database of synthetic, converted and replayed speech. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019, pp 1–24. https://doi.org/10.1016/j.csl.2020.101114, 1911.01601
132. Wang Z, Cui S, Kang X, Sun W, Li Z (2020) Densely connected convolutional network for audio spoofing detection. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp 1352–1360
133. Lin W (2015) An improved GMM-based clustering algorithm for efficient speaker identification. in: 2015 4th international conference on computer science and network technology ICCSNT), IEEE, pp 1490–1493. https://doi.org/10.1109/ICCSNT.2015.7491011
134. Wijethunga R, Matheesha D, Noman AA, De Silva K, Tissera M, Rupasinghe L (2020) Deepfake audio detection: a deep learning based solution for group conversations. In: 2020 2Nd international conference on advancements in computing (ICAC), IEEE, pp 192–197. https://doi.org/10.1109/ICAC51239.2020.9357161
135. Wu Z, Li H (2016) On the study of replay and voice conversion attacks to text-dependent speaker verification. Multimedia Tools and Applications. pp 5311–5327. https://doi.org/10.1007/s11042-015-3080-9
136. Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and countermeasures for speaker verification: a survey. Speech Comm 66:130–153. https://doi.org/10.1016/j.specom.2014.10.005
137. Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilçi C, Sahidullah M, Sizov A (2015) ASVSpoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2015, pp 2037–2041
138. Wu Z, Yamagishi J, Kinnunen T, Hanilçi C, Sahidullah M, Sizov A, Evans N, Todisco M, Delgado H (2017) ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. IEEE J Select Top Signal Process 11(4):588–604. https://doi.org/10.1109/JSTSP.2017.2671435
139. Wu Z, Yamagishi J, Kinnunen T, Hanilçi C, Sahidullah M, Sizov A, Evans N, Todisco M, Delgado H (2017) ASVSpoof: The automatic speaker verification spoofing and countermeasures challenge. IEEE J Select Top Signal Process 11(4):588–604. https://doi.org/10.1109/JSTSP.2017.2671435
140. Wu Z, Das RK, Yang J, Li H (2020) Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In: Interspeech 2020, ISCA, ISCA, pp 1101–1105. https://doi.org/10.21437/Interspeech.2020-1810
141. Yang J, Yang JY, Zhang D, Lu JF (2003) Feature fusion: Parallel strategy vs. serial strategy. Pattern Recogn 36(6):1369–1381. https://doi.org/10.1016/S0031-3203(02)00262-5
142. Yang J, Das RK, Li H (2019) Extended Constant-Q cepstral coefficients for detection of spoofing attacks. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings, APSIPA organization, pp 1024–1029. https://doi.org/10.23919/APSIPA.2018.8659537
143. Ye Y, Lao L, Yan D, Lin L (2019) Detection of replay attack based on normalized constant q cepstral feature. In: 2019 IEEE 4Th international conference on cloud computing and big data analysis (ICCCBDA), IEEE, pp 407–411. https://doi.org/10.1109/ICCCBDA.2019.8725688
144. Zeinali H, Stafylakis T, Athanasopoulou G, Rohdin J, Gkinis I, Burget L, Ěrnocký J (2019) Detecting spoofing attacks using VGG and SINCNET: But-omilia submission to AsvSpoof 2019 challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 1073–1077. https://doi.org/10.21437/Interspeech.2019-2892, 1907.12908
145. Zhang C, Cheng J, Gu Y, Wang H, Ma J, Wang S, Xiao J (2020) Improving replay detection system with channel consistency DenseNeXt for the ASVspoof 2019 challenge. In: Interspeech 2020, ISCA, ISCA, pp 4596–4600. https://doi.org/10.21437/Interspeech.2020-1044



Choon Beng Tan received his BCompSc. and MSc. degrees from Universiti Malaysia Sabah (UMS) in 2016 and 2018. He is now doing his PhD study in Computer Science in Universiti Malaysia Sabah (UMS). His recent work includes malware classification using machine learning and ensemble techniques, and cloud data integrity scheme; he is now working on voice presentation attack detection. His research interests include information security, cloud computing, and voice biometric security.

Mohd Hanafi Ahmad Hijazi is an Associate Professor of Computer Science at the Faculty of Computing and Informatics, Universiti Malaysia Sabah in Malaysia. His research work addresses the challenges in knowledge discovery and data mining to identify patterns for prediction on structured and/ or unstructured data; his particular application domains are medical image analysis and understanding and sentiment analysis on social media data. He has authored/ co-authored more than 50 journals/ book chapters and conference papers, most of which are indexed by Scopus and ISI Web of Science. He also served on the program and organizing committees of numerous national and international conferences. He is the leader of Data Technologies and Applications research group at the faculty.

Norazlina Khamis received her BIT (Hons) from University of Malaya, in 1999; her MSc degree in Realtime Software Engineering from Universiti Teknologi Malaysia in 2001; and her PhD from Universiti Kebangsaan Malaysia in 2012. She is currently attached as a senior lecturer with Universiti Malaysia Sabah. Her research interests include software engineering, intelligent software engineering, Internet of Things and Software Quality. She involves in several research related to ICT in disaster management system. She is a Fellow in Natural Disaster Research Centre, UMS. Currently she is working with several project related with Internet of Things.

Dr Puteri Nor Ellyza binti Nohuddin received her BSc. in Computer Science from University of Missouri-Columbia, USA and her MSc IT from Universiti Teknologi MARA. In 2012, she was awarded her Ph.D. in Computer Science from the University of Liverpool, UK. Puteri joins Institute of IR4.0 (IIR4.0), Universiti Kebangsaan Malaysia as a Research Fellow in July 2015. Prior to coming to IIR4.0, she was a lecturer at the Universiti Pertahanan Nasional Malaysia, Kuala Lumpur. Prior to her academic career, she worked with several conglomerates such as ExxonMobil, Sime Darby, Shell IT and Malaysian Resources Corporation Berhad as System Analyst. Puteri's teaching interests include Programming, Database systems and Data mining. Her primary research interests are in the field of Big Data, Data Mining and Knowledge Engineering. Specifically, she is interested in Time Series Clustering, Trend mining, Tacit Knowledge, and Social Network Analysis.

Dr Zuraini Zainol is a senior lecturer at the Department of Computer Science, National Defence University of Malaysia (NDUM). Prior to coming to NDUM, she was a lecturer at the Akademi Tentera Malaysia (ATMA), Kuala Lumpur and Cybernetics International College of Technology (CICM), Kuala Lumpur. Zuraini received her BSc. in Computer Science from Universiti Sains Malaysia (USM) and her MSc in Computer Science from Universiti Putra Malaysia (UPM). She completed her Ph.D. in Informatics from the University of Reading (UoR), UK in 2013. Zuraini's teaching interests includes Programming, Data Analytics, Data Mining and Database Systems. Her field of research interests are Text Analytics, Data Mining, Knowledge Management, and Artificial Intelligence. She has published over 50 journal and conference publications based on her research and managed several research grants to completion.

Frans Coenen has a general background in AI, and has been working in the field of data mining and Knowledge Discovery in Data (KDD) for the last fifteen years. He is interested in the application of the techniques of data mining and Knowledge Discovery in Data to unusual data sets, such as: (i) graphs and social networks, (ii) time series, (iii) free text of all kinds, (iv) 2D and 3D images, particularly medical images, and (v) video data. He is also interested in data mining over encrypted data. He currently leads a small research group working on many aspects of data mining and KDD. He has some 390 refereed publications on KDD and AI related research, and has been on the programme committees for many KDD conferences and related events. Frans Coenen is currently professor within the Department of Computer Science at the University of Liverpool where he is the director for the University of Liverpool Doctoral Network in AI for Future Digital Health.

Abdullah Gani received the B.Phil. and M.Sc. degrees in information management from University of Hull, U.K., and the Ph.D. degree in computer science from University of Sheffield, U.K. Prior to his degree studies, he acquired the Teaching Certificate from the Kinta Teaching College, Ipoh, and the Diploma degree in computer science from ITM. He has vast teaching experience due to having worked in a number of educational institutions locally and abroad-schools, the Malay Women Teaching College, Melaka, Ministry of Education; the Rotterham College of Technology and Art, Rotterham, U.K.; and the University of Sheffield, U.K. He is currently a Professor with the Dean Faculty of Computing and Informatics, Universiti Malaysia Sabah, and also an Honory Professor with the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He is also working on mobile cloud computing, big data, and the IoT. He is also the Dean of Faculty of Computing and Informatics, Universiti Malaysia Sabah. His interest in research kicked off In 1983 when he was chosen to attend the 3-Month Scientific Research Course in RECSAM, Ministry of Education, Malaysia. His current research interests include self-organized systems, machine learning, reinforcement learning, and wireless related networks. He was elected as a Fellow of the Academy of Science Malaysia (FASc) for Engineering and Computer Science Discipline.