Artificial Intelligence Review
https://doi.org/10.1007/s10462-023-10539-8

Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures

Awais Khan[1] · **Khalid Mahmood Malik**[1] · **James Ryan**[1] · **Mikul Saravanan**[1]

Accepted: 13 June 2023

Abstract
With the advent of automated speaker verification (ASV) systems comes an equal and opposite development: malicious actors may seek to use voice spoofing attacks to fool those same systems. Various counter measures have been proposed to detect these spoofing attacks, but current offerings in this arena fall short of a unified and generalized approach applicable in real-world scenarios. For this reason, defensive measures for ASV systems produced in the last 6-7 years need to be classified, and qualitative and quantitative comparisons of state-of-the-art (SOTA) counter measures should be performed to assess the effectiveness of these systems against real-world attacks. Hence, in this work, we conduct a review of the literature on spoofing detection using hand-crafted features, deep learning, and end-to-end spoofing countermeasure solutions to detect logical access attacks, such as speech synthesis and voice conversion, and physical access attacks, i.e., replay attacks. Additionally, we review integrated and unified solutions to voice spoofing evaluation and speaker verification, and adversarial and anti-forensic attacks on both voice counter measures and ASV systems. In an extensive experimental analysis, the limitations and challenges of existing spoofing counter measures are presented, the performance of these counter measures on several datasets is reported, and cross-corpus evaluations are performed, something that is nearly absent in the existing literature, in order to assess the generalizability of existing solutions. For the experiments, we employ the ASVspoof2019, ASVspoof2021, and VSDC datasets along with GMM, SVM, CNN, and CNN-GRU classifiers. For reproducibility of the results, the code of the testbed can be found at our GitHub Repository (https://github.com/smileslab/Comparative-Analysis-Voice-Spoofing).

Keywords Voice spoofing detection · Voice spoofing counter measures · ASVspoof · Deepfake speech detection · Speech synthesis · Deepfake audio detection

1 Introduction

Automatic speaker verification systems (ASVs) are now in wide use for authentication, e.g., for over-the-phone banking, and are becoming an increasingly important biometric solution in the current climate of the COVID-19 pandemic to limit the spread of

Extended author information available on the last page of the article

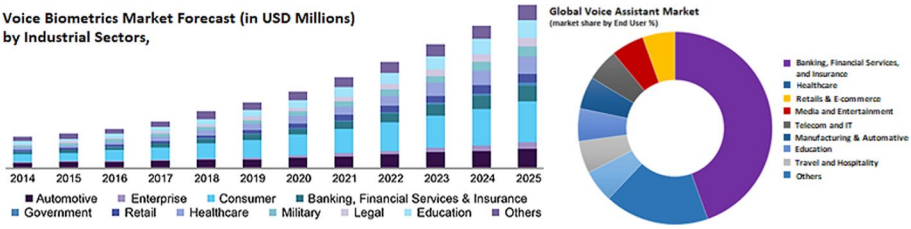

Fig. 1 The overall trend analysis of the voice biometrics and voice assistant based systems

disease. In addition, ASV systems are gaining traction in voice biometric systems and IoT devices as described in Malik et al. (2020); Javed et al. (2021). Voice biometrics, in particular, are being more widely used across a wide range of industries, including automotive, finance, education, and healthcare. Figure 1 shows the global trend and projected adoption of speech biometrics and voice assistants. However, although ASVs can be used to validate an identity they are prone to spoofing attacks. The statistical information presented in Fig. 1, pertains to the voice biometrics market forecast and the global voice assistant application market. To gather this data, we conducted extensive research using relevant keywords such as "voice biometrics market forecast" and "Global voice assistant application market." The information was obtained from reputable sources, including grand view research grandviewresearch (2023) and expert market research reports Research (2023).

Spoofing attacks on ASVs can be grouped into Physical Access (PA) attacks and Logical Access (LA) attacks as discussed in Balamurali et al. (2019). An example of a PA attack is a replay attack, where a hacker records the victim's voice and replays it to an ASV system in order to gain access to the asset the ASV system is protecting as shown in Malik et al. (2020). LA attacks, conversely, are comprised of artificial, i.e., machine-generated, cloned samples. LA attacks may consist of text-to-speech (TTS) synthesis and VC attacks, among others. Again, the objective is to generate realistic voice samples of a target speaker in order to compromise the security of an ASV system, and thus gain access to an asset. A more recent threat to ASVs and spoofing counter measures is adversarial machine learning Szegedy et al. (2014), where machine learning algorithms may be deceived by injecting minor perturbation into the audio sample. According to Szegedy et al. (2014), deep learning model predictions may be easily manipulated by extremely small perturbations in the data.

Several counter measures have been proposed to defeat voice spoofing attacks, and these are often comprised of two parts: the first one (front-end) is a feature representation scheme for the input speech signal, and the second one (back-end) is a classifier to distinguish between bonafide and spoofed samples. The feature descriptor (front-end) should be capable of effectively capturing the traits of the dynamic vocal tracts of a bonafide speaker. Similarly, the back-end classifier should be able to better learn the distinct traits of bonafide and spoofed speech samples in order to accurately discriminate against spoofed speech. In contrast to the traditional approach of front- and back-end solutions, in the past few years the research community has focused on deep learning and end-to-end solutions to combat voice spoofing attacks. Recent efforts toward in finding a unified solution to the problem of speaker verification, where a single countermeasure may be applied to several attacks, have been attempted, albeit with marginal success. Although these countermeasure solutions are

only beginning to be explored, there is a marked need for a unified solution as the way forward in ASV anti-spoofing techniques Javed et al. (2022).

The survey proposed in this paper investigates significant contributions to the ASV system development chain. More specifically, the main contributions of our work are:

- In the first work from the perspective of generalizability, and based on a thorough literature analysis of relevant attacks and counter measures, we present an in-depth overview and experimental analysis of the SOTA, development, challenges, and future direction of voice anti-spoofing research.
- We present the key limitations of the SOTA counter measures, developed by in-field researchers, and present the details of publicly available datasets used to benchmark the performance of voice anti-spoofing solutions.
- We address the need for cross-corpus evaluation and generalization by evaluating the performance of the featured counter measures, in terms of minimum tandem detection cost function (min t-DCF) and equal error rate (EER), on three large-scale, publicly available, and diverse datasets, and four different machine learning and deep learning classifiers.
- We make available our GitHub repository,[1] which contains the code and configuration requirements needed to perform the comparative analysis and ensure the reproduciblity of the results.

The rest of the paper is structured as follows: Sect. 2 describes the literature selection criteria and the existing surveys that have been published to date. Section 3 provides a detailed introduction to the possible attacks on ASV systems. Section 4 presents the SOTA counter measures developed to encounter voice spoof detection. The datasets developed to evaluate the performance of the counter measures presented in Sect. 5. Performance evaluation metrics are presented in Sect. 6. Sections 7 and 8 detail the experimental evaluation of counter measures using three different datasets, as well as their internal and cross-corpus evaluation. Sections 9 and 10 present experimental observations and limitations, conclusion and future work.

2 Existing surveys on voice anti-spoofing techniques

2.1 Literature selection criteria

In this survey, we review existing research papers that focus on techniques for detecting and countering audio spoofing attacks. A detailed description of the approach and protocols employed for the review is given in Table 1. We also display trends in spoofing attacks and counter measures year over year by examining Google Scholar papers released in the previous 6 years (2015–2023). Figure 2 depicts the course of the published study.

[1] https://github.com/smileslab/Comparative-Analysis-Voice-Spoofing

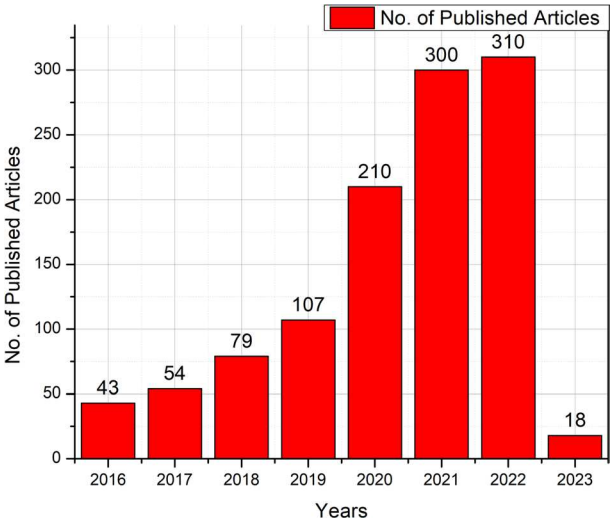

Fig. 2 Number of papers in the area of voice spoofing attacks and counter measures by year of publication

Table 1 Literature collection protocol

Preparation protocol	Description
Purpose	To identify current state-of-art voice spoofing attacks and counter measures. To critically compare both popular approaches and the features used in voice spoofing counter measures. To investigate open challenges in the domain of voice spoofing counter measures, and paths forward.
Sources	Google Scholar, IEEE explore, Springer Link
Query	Queries were used on the data sources above for collection of sources: Audio Spoof Countermeasure/ Spoofed Audio Detection/ Audio Synthesis Detection/ Audio Replay Detection/ Audio Spoofing counter measures/ Automatic Speaker Verification/ Secure ASV System/ ASV spoofing/ Presentation Attacks/ VC/ Adversarial Attacks on ASV/ Voice Replay Attacks/ Anti Spoofing counter measures/ Voice Spoofing Detection.
Method	Literature was categorized as follows: Audio spoofing detection and counter measures based on traditional and deep-learned methods, as well as handcrafted features. Existing voice spoofing attacks, along with adversarial attacks and their taxonomy. An extensive examination, testing and comparison of existing voice spoofing counter measures using single and cross corpus evaluation. Detailed discussion of the research gap, issues, limitation, and future trends in voice spoofing detection and counter measures.
Size	A total of 150 papers were retrieved using the above query until the search was discontinued on 01-10-2023. We selected literature that was relevant to the subject of audio spoof detection and excluded those that were not relevant to this subject or were white papers/articles.
Inclusions and exclusions	Preference was given to peer-reviewed journal papers and conference proceedings articles. In addition, articles from the archive literature were also taken into account.

2.2 Analysis of the six existing surveys

The surveys on ASV systems and spoof detection techniques conducted to date have primarily been focused on specific spoofing attacks and the counter measures for them. To the best of our knowledge, six articles Wu et al. (2015); Patil and Kamble (2018); Sahidullah et al. (2019); Kamble et al. (2020); Tan et al. (2021); Mittal and Dua (2021) have been presented as surveys, and two studies Sahidullah et al. (2016) and Font et al. (2017) conducted a comparative analysis of voice spoofing counter measures.

In a survey of voice spoofing counter measures, Wu et al. (2015, 2015a, 2015b), provided a comprehensive taxonomy of voice spoofing attacks known at the time, along with ASV system vulnerabilities. However, the age of this review means the advanced spoofing attacks that have grown prevalent in ASV systems were not addressed. In addition, it is focused on a dedicated countermeasure for each sort of attack, rather than exploring a more unified solution. Following this, Sahidullah et al., Sahidullah et al. (2016) and Font et al. (2017) published comparative analyses of the spoofing counter measures that focused solely on replay attacks. According to their work, published during 2017–2018, replay attacks attracted the most attention in the wake of the ASVspoof 2017 challenge. The results of Sahidullah et al. (2016), using the ASVspoof2015 dataset, demonstrated that an ensemble of acoustic features, together with machine learning models, produced more accurate results than individual classifiers. Kamble et al. Kamble et al. (2020) also published a work that covered a subset of the specific speech corpora along with the evaluation measures in this field.

Font et al. (2017) performed a comparative analysis of nine different counter measures with cross-corpus evaluation of the counter measures. However, the authors only discussed the performance of the counter measures against replay attacks due to the nature of the dataset used for the study. Although Sahidullah et al. (2016) study and Font et al. (2017) paper each provided a comprehensive performance evaluation of the replay attack, other types of spoofing attacks that could severely disrupt an ASV system were ignored. Sahidullah et al. surveyed four different types of spoofing attacks, spoofing procedures, and counter measures the next year, in 2019 Sahidullah et al. (2019), summarized some spoofing challenges and presented counter measures. However, a comprehensive analysis of all aspects of the ASV system, including end-to-end counter measures, cross-corpus evaluation using modern attack-oriented datasets, and a performance evaluation of both unified and integrated ASV solutions, was still lacking. For instance, Sahidullah et al. (2016) and Font et al. (2017) focused only on front-end countermeasure design, and concentrated on the SS, VC, replay, and mimicking attack types, alone where speech corpora, protocols, classifier, and evaluation metrics all contributed equally to the construction of an ASV system. These were absent from the Sahidullah et al. (2019) study. Lastly, the existing work also lacks the recent encoded and compressed logical and physical attacks presented in the ASVspoof2021 dataset.

2.3 Lessons from prior surveys

With the notable exception of Tan et al. (2021), the articles by and large did not present a detailed taxonomy of modern voice presentation attack detection (PAD) methods, where the work was structured based on identified elements. This survey Tan et al. (2021), along with Mittal and Dua (2021) broadened the classification of relevant work and built on the taxonomy from the most recent work on voice PAD. The significant contribution of these

Table 2 Comparison of the existing survey and review papers

Paper	Presentation attacks			counter measures			I-ASV[a]	EA[b]	CC[c]
	Replay	VC/TTS	Adversarial	HC[d]	DL[e]	E2E[f]			
Wu et al. (2015)	✓	✓	×	✓	✓	✓	×	×	×
Sahidullah et al. (2016)	✓	×	×	✓	×	×	×	✓	×
Font et al. (2017)	✓	×	×	✓	×	×	×	✓	✓
Patil and Kamble (2018)	✓	×	×	✓	×	×	×	×	×
Sahidullah et al. (2019)	✓	✓	×	✓	✓	✓	×	×	×
Tan et al. (2021)	✓	✓	×	✓	✓	✓	×	×	×
Mittal and Dua (2021)	✓	✓	×	✓	✓	✓	×	×	×
Kamble et al. (2020)	✓	✓	×	✓	✓	✓	×	×	×
[Ours]	✓	✓	✓	✓	✓	✓	✓	✓	✓

[a] I-ASV refers to Integrated ASV solutions

[b] EA denotes the experimental analysis

[c] CC denotes the cross corpus evaluation of the counter measures

[d] HC refers to Hand Crafted Solutions

[e] DL refers to Deep Learning

[f] E2E denotes the end-to-end solutions

articles was the presentation of trends and analyses of voice PAD that were absent in the other survey articles. Although Mittal and Dua (2021) described the computation mechanisms of traditional and modern speech feature extraction, as well as datasets and a combination of various front- and back-end techniques, none of these articles performed a fair experimental analysis of the existing SOTA counter measures. In addition, existing reviews and survey articles also lack the cross-corpus evaluation which would show the generalizability of existing solutions. A comparison of the existing reviews and surveys is presented in Table 2.

In addition, existing spoofing counter measures have employed diverse features and classifiers and, typically, have had their performance evaluated on only one of the several extant datasets, e.g., ASVspoof 2015, ASVspoof 2017, and ASVspoof 2019, using different metrics. With no standardization, it is difficult to declare a single countermeasure as the method that works best. Thus, there exists a need to provide a thorough analysis of existing counter measures in order to show which method is the best fit for a certain scenario. To the best of our knowledge, no comparative analysis work on multiple voice spoofing attacks, including single and multi-order attacks, has ever been presented. Moreover, existing studies have ignored the important aspect of cross-corpora evaluation, which is crucial to the evaluation of the generalized nature of the countermeasure. In consequence, the focus of this study is on addressing the existing issues and limitations of the developed anti-spoofing systems, as well as presenting a comparative and cross corpus examination of various spoofing counter measures.

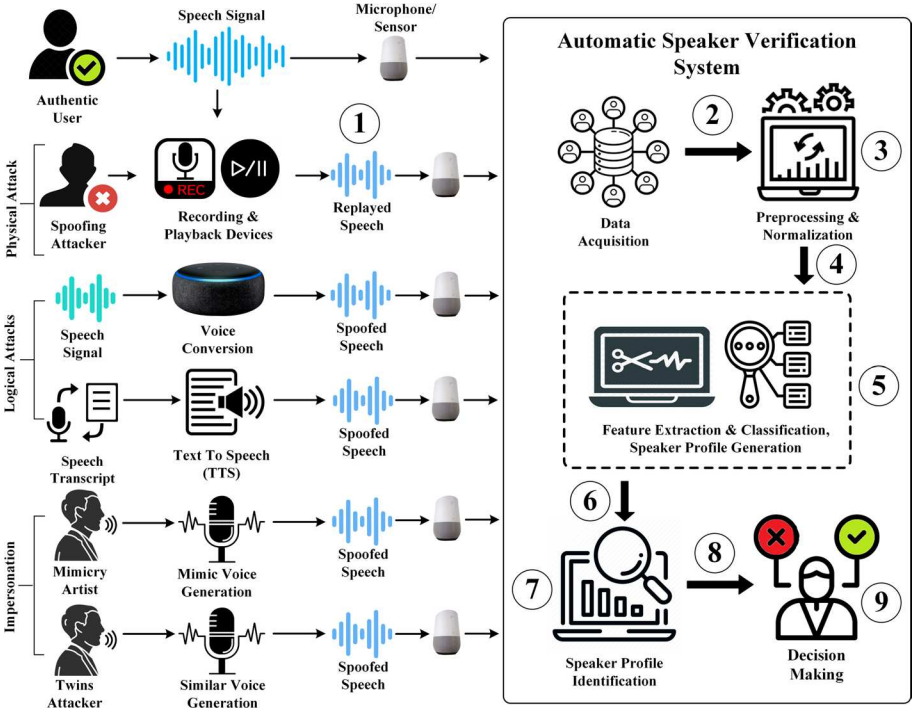

Fig. 3 Existing threats to automatic speaker verification systems

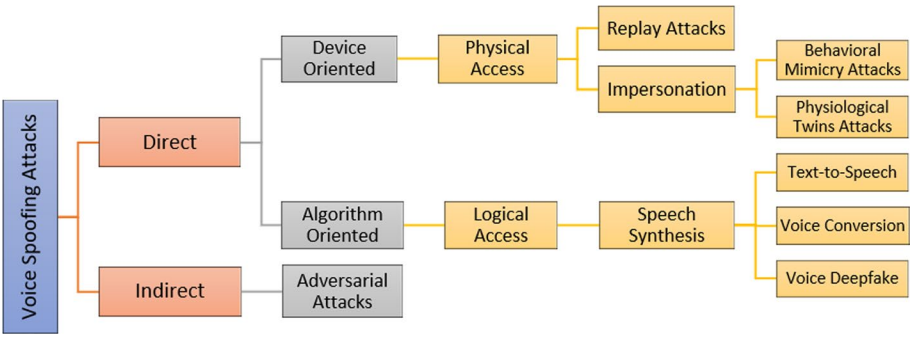

Fig. 4 Taxonomy of the voice spoofing attacks

3 Voice spoofing attacks on ASV systems

ASV systems, are vulnerable to a variety of direct voice spoofing attacks, i.e., physical-access (PA), logical-access (LA) attacks, and indirect attacks, i.e., adversarial attacks, where the audio signal remains unchanged but the attacker modifies the signal's properties during ASV processing. These attacks on ASV systems are graphically illustrated in Fig. 3 and the taxonomy of the attacks are presented in Fig. 4.

3.1 Physical attacks

PA attacks occur when the spoofed samples are applied as an input to the ASV system through the sensor, such as a microphone, with replay attacks as the prime example, which occur when the audio of an authorized user is recorded and played back in order to deceive an authentication system. As demonstrated in our prior work Baumann et al. (2021), replay attacks may be single order, where a recorded sample is directly played back to an ASV system, or multi-order, where an attack is replayed to one or more recording devices and the secondary or higher recording is played to the ASV. Another PA, the impersonation attack, occurs at the microphone level, where an imposter changes how they talk to mimic the speech characteristics of a legitimate user, which has been shown to be successful if the imposter's natural voice has similar features.

3.2 Logical attacks

LA attacks occur when the audio samples bypass the sensor and are injected directly into the speaker verification system. Spoofing attacks based on logical access include VC and SS. VC uses an imposter's natural voice to generate artificial speech in order to match the targeted speaker's voice. Attacks are usually created by other models to fool a specific ASV system. Machine learning has allowed the mapping of speech features between speakers to be accurate, and is now computationally efficient enough to make ASV systems vulnerable to these types of attacks. However, these attacks can still be detected because they are not a perfect match to genuine audio. SS attacks, also known as deepfake attacks, are similar to VC but use text as an input to the model in order to generate a voice clip similar to a targeted speaker in an effort to fool an ASV system. Models that generate accurate speech features can be trained using a small data set of recorded audio.

3.3 Adversarial attacks

Adversarial attacks make up the most recent cyber threats to ASVs. An adversarial attack is a malicious attempt on the machine learning model itself, to intentionally cause misclassifications, usually using a slight perturbation of the original input. Adversarial attacks may also be attempted as described in points 2–9 in Fig. 3. This attack has had an influence on a variety of intelligent domains involving image, audio, and video sample identification, as well as essential security applications such as intrusion detection and malware detection. Machine learning models are extremely susceptible to adversarial attacks. According to Jati et al. (2021), ASV systems are prone to adversarial attacks, which may reduce the accuracy of such systems by up to 94%. Attacks can range from simple Gaussian noise to more advanced attacks created in a targeted white-box setting. These perturbations, imperceptible to humans, may cause audio classification and ASV systems to fail completely.

Table 3 A comparison of existing SOTA papers

Paper	Datasets	Presentation attacks			counter measures			
		PA	LA	DF[a]	HC/C[b]	DL[c]	E2E[d]	CC[e]
Paul et al. (2015)	ASVspoof15	X	✓	X	✓	X	X	X
Kinnunen et al. (2017a)	ASVspoof17	✓	X	X	✓	X	X	X
Ji et al. (2017)	ASVspoof17	✓	X	X	✓	X	X	X
Balamurali et al. (2019)	ASVspoof17	✓	X	X	✓	✓	X	X
Tapkir and Patil (2018)	ASVspoof17	✓	X	X	✓	X	X	X
Novoselov et al. (2016)	ASVspoof15	X	✓	X	✓	X	X	X
Tapkir et al. (2018)	ASVspoof17	✓	X	X	✓	X	X	X
Cai et al. (2017)	ASVspoof17	✓	X	X	✓	✓	X	X
Yang and Das (2019)	ASVspoof17v2.0	✓	X	X	✓	X	X	X
Javed et al. (2022)	ASVspoof19 & VSDC	✓	X	X	✓	X	X	X
Hassan and Javed (2021)	ASVspoof19-LA	X	✓	X	✓	X	X	X
Arif et al. (2021)	ASVspoof19-LA	X	✓	X	✓	✓	X	X
Gunendradasan et al. (2018)	ASVspoof17	✓	X	X	✓	X	X	X
Witkowski et al. (2017)	ASVspoof17	✓	X	X	✓	X	X	X
Gunendradasan et al. (2019)	ASVspoof17	✓	X	X	✓	X	X	X
Saranya et al. (2018)	ASVspoof17	✓	X	X	✓	X	X	X
Yang and Das (2019)	ASVspoof15 & 19	X	✓	X	✓	X	X	X
Chen et al. (2015)	ASVspoof15	X	✓	X	X	✓	X	X
Nagarsheth et al. (2017)	ASVspoof17	✓	X	X	✓	✓	X	X
Lai et al. (2019)	ASVspoof17	✓	X	X	✓	✓	X	X
Huang and Pun (2020)	ASVspoof17	✓	X	X	✓	✓	X	X
Wu et al. (2020)	ASVspoof19	X	✓	X	✓	✓	X	X
Ma et al. (2021)	ASVspoof19	X	✓	X	X	✓	X	X
Tak et al. (2021)	ASVspoof19	X	✓	X	X	✓	X	X
Liu and Yang (2020)	ASVspoof19	✓	✓	X	✓	✓	X	X
Yang et al. (2019)	ASVspoof15 & 17v2.0	✓	✓	X	✓	✓	X	X
Suthokumar et al. (2019)	ASVspoof17v2.0	✓	X	X	✓	✓	X	X
Wang et al. (2019)	ASVspoof17v2.0	✓	X	X	✓	✓	X	X
Wu et al. (2018)	ASVspoof17	✓	X	X	✓	✓	X	X
Balamurali et al. (2019)	ASVspoof15 & 17v2.0	✓	✓	X	✓	✓	X	X
Dinkel et al. (2017)	BTAS16 & ASVspoof15	✓	X	X	X	✓	✓	X
Jung et al. (2021)	ASVspoof21	✓	✓	X	X	✓	✓	X
Tak et al. (2021)	ASVspoof19	✓	✓	X	X	✓	✓	X
Javed et al. (2021)	ASVspoof19	✓	✓	X	✓	✓	X	X
Rostami et al. (2021)	ASVspoof19	✓	✓	X	✓	✓	X	X
Lai et al. (2019)	ASVspoof19	✓	✓	X	✓	✓	X	X
Chen et al. (2021)	ASVspoof21	X	✓	✓	X	✓	X	X
Jung et al. (2022)	Voxceleb & ASVspoof19	✓	✓	X	X	✓	X	X
Zhang et al. (2022)	Voxceleb & ASVspoof19	✓	✓	X	X	✓	X	X
Liu et al. (2022)	ASVspoof19	✓	✓	X	X	✓	X	X
Teng et al. (2022)	ASVspoof19	✓	✓	X	X	✓	X	X
Jose et al. (2018)	chsc2011 & GTZAN	X	X	X	✓	X	X	X

Table 3 (continued)

Paper	Datasets	Presentation attacks			counter measures			
		PA	LA	DF[a]	HC/C[b]	DL[c]	E2E[d]	CC[e]
Kua et al. (2010)	NIST2001& & SRE	*X*	*X*	*X*	✓	*X*	*X*	*X*
Wu et al. (2017)	ASVspoof15	✓	✓	*X*	✓	✓	*X*	*X*
Rajan et al. (2013)	NIST 2010 Eval	*X*	*X*	*X*	*X*	✓	*X*	*X*
Malik et al. (2020)	VSDC & ASVspoof19	✓	✓	*X*	✓	✓	*X*	*X*
Chettri et al. (2020)	ASVspoof19-PA	✓	*X*	*X*	✓	✓	*X*	*X*
Chen et al. (2017)	ASVspoof19-PA	✓	*X*	*X*	*X*	✓	✓	*X*
Naika (2018)	VoxForge	✓	*X*	*X*	*X*	✓	✓	*X*

DF denotes the deepfake voice attacks, HC/C refers to Hand Crafted features and Conventional classifiers i.e., GMM, UBM etc. DL refers to Deep Learning, E2E denotes the end-to-end solutions and CC denotes the cross corpus evaluation of the counter measures

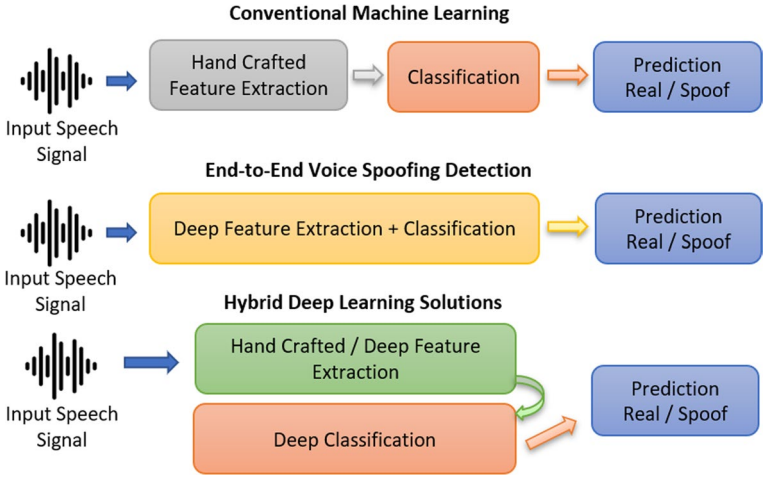

Fig. 5 Existing approaches of voice spoofing counter measures to counter voice spoofing attacks

4 Voice spoofing counter measures, taxonomy, and analysis

This section provides a detailed analysis of existing voice spoofing counter measures. Though not intended to be a comprehensive record of all anti-spoofing methods, this carefully curated list is an attempt to summarize the solutions which may be considered SOTA at the time of this writing. These solutions may be loosely grouped into five classifications, as seen in the following subsections. The existing approaches used in counter measures and their taxonomy are depicted in Figs. 5 and 6 offers a thorough classification of the counter measures discussed in this study. While a comparison of the existing hand crafted, deep learning and End-to-end solution is presented in Table 3.

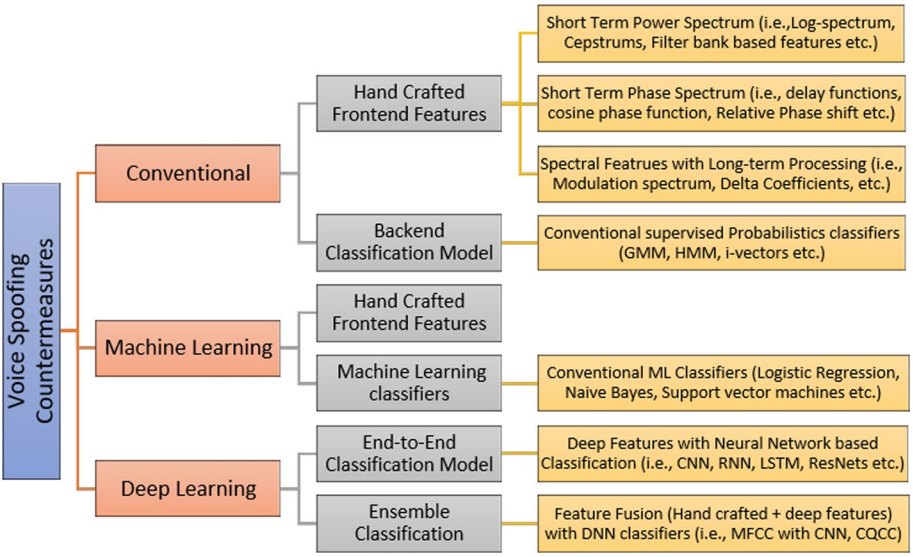

Fig. 6 Taxonomy of the voice counter measures

4.1 Handcrafted feature-based solutions

Hand-crafted features refer to a set of speech characteristics that are manually selected and extracted by experts in the field of speech processing and are used to identify voice spoofing. Because of their ability to capture the unique speech characteristics that are indicative of spoofing attempts, hand-crafted features have been frequently employed in voice spoofing detection. For instance, in a study Paul et al. (2015), the author derived novel acoustic features from the frequency-warping and block transformation of filter bank log energies to classify bona fide and spoofed speech samples. Although the proposed features achieved a 0.99% classification accuracy on the ASVspoof 2015 corpus development section, they were only applicable when the spoofing attacks were known in advance. Aside from this, the performance of the proposed approach was not evaluated across corpora.

In Sahidullah et al. (2016), Sahidullah et al. presented the first comparative evaluation of six counter measures and their integration with ASV systems using the ASVspoof 2015 dataset. The counter measures contain MFCCs, IMFCCs, LFCCs, CFCCs, CQCCs, and GFCCs as front-end cepstral features, which are then coupled with GMM-UBM and i-Vector classifiers for back-end speaker classification. According to their results, the countermeasure using only CQCC features and the countermeasure with the fusion of all six counter measures have the greatest potential to detect spoofing attacks. More significantly, the fusion of counter measures and CQCCs achieved the lowest EER of 0.02%. The results of Sahidullah et al. (2016) demonstrate that an ensemble of acoustic features, together with machine learning models, can produce more accurate results than individual classifiers. Although the significance of CQCC features is demonstrated in this study, the presented system was not evaluated against other forms of spoofing attacks, e.g., replay, deepfake, adversarial, or others. In Kinnunen et al. (2017a), mainly followed this line of thinking, utilizing front-end features, i.e., CQCC, MFCC, RFCC, and HPCC, etc., with a baseline

GMM-UBM classifier. The performance of these counter measures was evaluated against the ASVspoof2017 dataset, where the proposed systems achieved an EER of 31.5% for genuine vs. replay and 1.8% EER for genuine vs. zero-effort imposters. When these handcrafted feature results were compared to those from the prior challenge (ASVspoof2015), the detection of replay attacks was found to be more complex than detecting SS and VC spoofing attacks.

Following the success of feature fusion in spoofing detection, in Ji et al. (2017), the author presents an ensemble of the acoustic features of CQCCs, along with other classical features, i.e., MFCCs and Perceptual Linear Predictive (PLP) features. In parallel, the author proposes an ensemble classifier set that contains numerous GMMs, i.e., Gradient Super Vector-Boosting Decision Tree (GSV-GBDT) and GSV-Random Forest (GSV-RF) classifiers. The experimental results show that the presented ensemble system significantly outperforms the baseline GSV-SVM system. Specifically, using a baseline GSV-SVM classifier, this approach achieves an EER of 10.4% on CQCC features, 27.4% on MFCC features, and 37.0% on PLP features. In contrast, a minimal EER of 9.5% is achieved by employing the ensemble model. Similarly to Paul et al. (2015), the biggest limitation of Ji et al. (2017) is that it has only been tested against a single replay attack. Other types of spoofing attacks, i.e., SS, VC, and the most current attacks facing ASV systems, i.e., deepfake and adversarial attacks, are not discussed, and there is no cross-corpus examination of the techniques.

In Balamurali et al. (2019), the authors examine the robust audio features, comprised of handcrafted and auto-encoder-based learned features, to identify replay spoofing attacks. The handcrafted features employed in this study are CQCCs, LPCCs, IMFCCs, Rectangular Filter Cepstral Coefficients (RFCCs), Sub-band Centroid Frequency Coefficients (SCFCs), and Sub-band Centroid Magnitude Coefficients (SCMC), as well as spectrogram features. Instead of using only hand crafted features with a back-end classifier the authors use an auto-encoder to learn a dense representation of all of the features. Later, a conventional GMM, along with a Universal Background Model (UBM), is used as the baseline system to examine the performance of the handcrafted and encoder-based features. The integrated fused models, based on existing audio and machine-learned features, achieve comparable results, with an EER of 12.0%. In particular, the handcrafted CQCC features outperform all other features, with an EER of 17.5%. The best encoder-based feature observed is a spectrogram with the minimum EER of 20.2%. Although the coupling of handcrafted features with an auto-encoder-based system surpasses SOTA PA systems, the presented system is only tested against replay attacks. In the presence of other types of spoofing attacks, e.g., VC and adversarial attacks, the performance of the system tends to vary.

Rather than using standard standalone short-term power spectrum coefficients, e.g., MFCCs, LFCCs, or CQCCs, to identify spoofing attacks, the authors of Tapkir and Patil (2018) introduced phase-based Teager Energy Operator (TEO) features. It is discovered that the TEO phase features give information that is complementary to the information provided by the more commonly used CQCC, MFCC, and LFCC feature sets. Although the TEO phase features are unable to perform well alone, fusion with traditional features enhances the accuracy of the spoofing detection system. The results demonstrate that the standalone spoof detection systems, developed with the TEO phase, MFCC, and LFCC, achieved an EER of 31.34, 34.02, and 16.80%, respectively, whereas, when the TEO phase feature set was fused with the CQCC, MFCC, and LFCC feature sets, the EER was lowered by 0.18%, 2.74%, and 1.41%, respectively. This improvement in system performance shows the influence of TEO phase information on the spoof detection system. However, the

provided solution is only stated to be resistant to replay spoofing attacks, and there is no cross-corpus validation.

In accordance with Tapkir and Patil (2018), the authors of Novoselov et al. (2016) present phase spectrum and multi-resolution wavelet features, in addition to the commonly used front-end MFCC features. This study combines MFCCs with Mel-Frequency Principal Coefficients (MFPCs), CosPhase Principal Coefficients (CosPhasePCs), and Mel Wavelet Packet Coefficients (MWPCs) to provide a reliable and robust defense against spoofing attacks. The experimental results on ASVspoof2015 dataset indicate that applying principal component analysis (PCA) to MFCCs results in a considerable EER improvement over MFPC features for all spoofing methods. However, the MFCC features prove inferior in comparison with other front-end features. In contrast, the implementation of CosPhasePC modestly decreases EER in comparison to the MFPC features. Although the multi-resolution wavelet transform features outperform the SOTA features, achieving 0.05% EER for all known attacks, the proposed framework has not been tested and reported cross-corpus or against unknown attacks.

Accordingly, following the significant performance of phase-oriented features, the author sheds light on the importance of acoustic front-end features and introduces a novel detection mechanism by modeling replayed speech as a convolution of original speech Tapkir et al. (2018). Also in Tapkir et al. (2018), the author proposes a novel feature set, Magnitude-based Spectral Root Cepstral Coefficients (MSRCC) and Phase-based Spectral Root Cepstral Coefficients (PSRCC), which outperforms the baseline system (CQCC) and provides a 29.18% EER on the evaluation set of the ASVspoof 2017 challenge database. With the GMM back-end classifier, the front-end features MSRCC and PSRCC respectively produce 18.61% and 24.35% EER. Conversely, with convolutional neural network (CNN) back-end classifiers, MSRCC and PSRCC obtain 24.50 and 26.81% EER, respectively. In addition, the score-level fusion of MSRCC and PSRCC results in 10.65 and 17.76% EER using GMM and CNN classifiers, respectively. These findings suggest that the proposed feature sets of MSRCC and PSRCC capture complementary information. However, once again a cross-corpus examination was not reported.

The research in Cai et al. (2017) shows the effectiveness of augmenting genuine training data in the simulation of replay spoofing attacks. The author presents replay spoofing countermeasure systems that improve the CQCC-GMM baseline with score level fusion. Instead of using CQCC, the author uses spectrograms as input to analyze end-to-end feature representations. Finally, the author replaces the baseline GMM classifier with a Fully-connected Deep Neural Network (FDNN) and a Bidirectional Long-Short-Term Memory neural network (Bi-LSTM). The results of the experiments show that this data augmentation technique can significantly increase the system's performance. In particular, the baseline CQCC-GMM model realizes an EER of 22.29%, while the DA-CQCC-GMM model obtains an EER of 19.18% and the fused system achieves an EER of 16.39%. Although the end-to-end FDNN and Bi-LSTM-based systems perform well for replay detection, other sorts of voice spoofing attacks such as deepfake and adversarial attacks may lead the system to perform inadequately. Furthermore, the provided system was not evaluated in the presence of unknown attacks, or across corpora.

Yang et al. Yang and Das (2019) developed a Low-Frequency Frame-wise Normalization (LFFN) technique to capture replay distortions. LFFN was combined with Constant-Q Transforms (CQT) to extract two features: Constant-Q Normalization Segmentation Coefficients (CQNSC) and Constant-Q Normalization Cepstral Coefficients (CQNCC). This approach performed well on the ASVspoof 2017 version 2.0 dataset, with an EER of 10.63% for CQNSC and 10.31% on the CQNCC features. Although a promising approach,

this method bears further investigation to determine the effectiveness of these features against unknown attack types, and across corpora, neither of which was presented.

The Acoustic Ternary Patterns-Gammatone Cepstral Coefficient (ATP-GTCC) feature was proposed in Javed et al. (2022) to help develop a lightweight model for single and multi-order replay attack detection. An SVM classifier was used to better capture the harmonic distortions found in multi-order replay samples, while ATP-GTCC was employed as a front-end feature. This model achieved an EER of 0.6 and 1% on the Voice Spoofing Detection Corpus (VSDC) and the ASVspoof 2019 dataset, respectively, which exceeded earlier SOTA techniques in terms of replay detection performance and efficiency. However, the given approach is limited to single-order and multi-order replay attacks, and has not been tested against other types of attacks, including VC, deepfake, and adversarial attacks.

In Hassan and Javed (2021), the authors use a feature fusion of GTCC, MFCC, Spectral Flux, and Spectral Centroid for input audio presentation. This countermeasure successfully detects multiple types of logical access attacks and classifies the cloning algorithms used to produce the synthetic speech. It achieves an EER of 3.05%, compared to baseline models that achieve an EER ranging between 5.06 and 9.57%. Although the presented system performed better against logical access attacks it is not tested in the presence of PA or adversarial attacks. The authors of Arif et al. (2021) present a voice spoofing countermeasure using ELTP-LFCC features and a Deep Bidirectional LSTM to combat TTS synthesis and converted voice samples in LA attacks. In this paper, ELTP is further fused with LFCC to better capture the characteristics of the vocal tract speech dynamics of both bonafide voice and cloning algorithm artifacts. On the diverse ASVspoof 2019-LA dataset, performance evaluation yields an EER of 0.74% and a min t-DCF of 0.008%. However, the presented system is only tested on one LA-dataset, and its performance against PA attacks is unknown. Other than the short-term spectral features, the frequency or amplitude modulation-based features were also explored for voice spoofing detection. Gunendradasan et al. Gunendradasan et al. (2018) used Spectral Centroid Deviation (SCD) features to develop a replay attack detection system. The Spectral Centroid Frequency (SCF) and Spectral Centroid Magnitude Coefficient (SCMC) features were extracted from the same front-end as SCD and used as complementary features to train a GMM classifier for replay attack detection. This method was evaluated on the ASVspoof2017 dataset and provided a 60% improvement with 9.20% EER as compared to the ASVspoof CQCC baseline model. However, performance of the system against other datasets or across corpora was not tested.

Some methods also exploit high-frequency components in order to capture the traits of bonafide and spoofed signals. The work in Witkowski et al. (2017) captured the high-frequency content by using the inverted-MFCC (IMFCC), LPCC, LPCCres, CQCC, MFCC, and Cepstrum features. This hybrid feature representation was then used to train a GMM for the classification of bonafide and spoofed samples. Due to the increased feature computation cost, this method was not suitable for local deployment on resource-constrained voice-controlled systems. This paper, Gunendradasan et al. (2019), also highlighted the idea of analyzing the high-frequency band for replay spoof detection. For this purpose, transmission line cochlea amplitude modulation and transmission line cochlea frequency modulation features were employed to train a GMM for replay attack detection. This method was evaluated on the ASVspoof2017 dataset and achieved an EER of 7.59%. It should be noted that this method performed only slightly better than the baseline model and has a high computational cost because the features take more than twice the amplitude frequency to modulate the signal, making this method unsuitable for resource-constrained ASVs. In the literature, non-voice segments have also been explored to capture the distortions of playback speech. Saranya et al., in Saranya et al. (2018), analyzed the channel and

reverberation information from non-voiced segments of the input audio for replay spoofing detection. A voice activity detector was used to determine the non-voiced segments, then a hybrid feature vector comprised of CQCC, MFCC, and Mel-Filter bank-Slope was employed to capture the remnant vocal tract information in the non-voiced segments of the input audio. These features were later used to train a GMM for the classification of bonafide and replayed audio samples. The performance of this method was evaluated on the ASVspoof2017 dataset, and showed an improvement of 37% with EER of 14.45% over the baseline method.

In the work of Yang and Das (2019), the authors introduced an inversion module to derive four features: ICQC, ICQCC, ICBC, and ICLBC. Two of them used conventional DCT, and the other two used overlapped block transforms. These features for synthetic speech detection were evaluated using the ASVspoof 2015, noisy ASVspoof 2015, and ASVspoof 2019 logical access datasets, and achieved an EER of 7.77%. Despite the fact that this system was tested on multiple datasets, due to advances in spoofing attacks, e.g., deepfakes, encoded LA and PA attacks, adversarial, or attacks where the artifacts of spoofing attempt are unknown, they may not perform well in a real-world scenario.

4.2 Discussion, challenges and the limitations of handcrafted-based counter measures

Although hand-crafted, feature-based solutions for voice spoofing detection have been widely studied they have some limitations and challenges that need to be addressed. One of the main limitations is that they are typically based on a fixed set of parameters and cannot adapt to different types of spoofing attacks. This indicates that they may not be able to capture the subtle differences between genuine and spoofed speech signals, especially in the case of unseen spoofing attacks, or mixed attacks, where the attack is not consistent across the entire sample. Another disadvantage is that they may not capture the complexity of the speech signal well because they are often based on basic signal processing techniques and are thus incapable of properly capturing nuances in the vocal signal. Another challenge is that hand-crafted, feature-based solutions may not be able to effectively handle variations in the speech signal, such as changes in speakers, accents, and recording environments. This may make it difficult to accurately classify incoming speech signals as genuine or spoofed in a practical scenario using the currently available datasets. Finally, it is unknown how existing systems will perform against complex spoofing attacks e.g., compression, distinct encoding artifacts, and frame-level partial spoofing distortions. Despite these limitations, hand-crafted features are still widely used in voice spoofing detection and have the advantage of being computationally efficient, easy to implement, and easy to interpret. They also provide a good representation of the speech signal and can be used in combination with other counter measures, such as deep learning techniques, to create robust and accurate systems. It is important to note that the use of hand-crafted features in voice spoofing detection is still an active area of research, and new features and techniques are continuously being proposed and evaluated. Therefore, it is important to keep up with the latest developments in the field, and to evaluate the performance of newly developed hand-crafted features for a given application.

4.3 Deep learning solutions

Recent years have witnessed a rise in the use of deep learning approaches to prevent audio spoofing attacks. In particular, deep learning classifiers, combined with handcrafted features, have been shown to improve the accuracy of voice spoofing detection systems. While deep learning models excel at learning complex representations from raw audio data, incorporating hand-crafted features offers an additional layer of expertise and domain-specific knowledge. This hybrid strategy has been shown to outperform pure, handcrafted solutions. In combination, the use of hand-crafted features acts as a form of regularization, guiding the deep learning models towards a solution that aligns with prior understanding and helps prevent overfitting to the training data. Consequently, deep learning solutions are becoming a widely adopted strategy in the field of voice spoofing detection. Some of the recent deep learning methods are discussed below in this section.

In Chen et al. (2015), a novel, simple model for detecting spoofing attacks on a speaker verification system was developed. The presented model was used to extract key features from audio samples and construct compact, abstract, and resilient deep data representations. A spoofing-discriminant network was used in the training of spoofing algorithms. The proposed network then computed the s-vector, which is the utterance level average of the final hidden layers. Finally, Mahalanobis distance, along with normalization, was used in conjunction with the computed s-vectors to detect spoofing attacks. This model achieved the 3rd position in the first spoofing detection challenge, ASVspoof 2015. In particular, the proposed model attained an overall minimum EER of 2.281%. Specifically, the presented system obtained an EER of 0.046% for known and 4.516% for unknown replay attacks. Although this system performed well in the ASVspoof 2015 competition, it has not been examined against recent attacks, replay attacks, or across corpora.

Existing methods also employed handcrafted features to train deep neural networks for spoofing detection. In Nagarsheth et al. (2017), high-frequency regions were analyzed by proposing the High Frequency Cepstral Coefficients (HFCC) feature and using a back-end DNN model to classify bonafide and spoofed (replayed) audio. When comparing HFCC features to CQCC features using a baseline GMM back-end, HFCC outperformed CQCC on the ASVspoof2017 development and evaluation sets. HFCC received an EER of 5.9% in the development set and 23.9% in the evaluation set, while CQCC received 11.0% EER in the development set and 24.7% in the evaluation set. Despite its superior performance in replay attack detection this system was not evaluated against VC and SS attacks.

An attentive filtering network system was proposed in Lai et al. (2019) to detect replay attacks using enhanced feature representations in both the frequency and time domains. This system obtained an EER of 6.09 and 8.54% on the development and evaluation sets of ASVspoof 2017, respectively, on a system comprised of two Attentive Filtering models (one using sigmoid and the other using softmax as their activation functions). This method achieved better results than the CQCC-GMM baseline model, which obtained an EER of 12.08 and 29.35%. However, the combination of feature extraction and the model's architecture required significant computational resources.

In Huang and Pun (2020), segment-based linear filter bank features, along with an attention-enhanced DenseNet-BiLSTM model, were proposed for replay attack detection. These features were extracted from the silent segments of the audio signal. Triangle filters were used to examine noise in the high frequency bands $3 - 8kHz$. The baseline system of CQCC-GMM obtained an EER of 2.36% on the development set and an EER of 8.42% on the evaluation set of ASVspoof2017, while features of this method, when used with a

GMM, obtained an EER of 1.8 and 7.92% on the development and evaluation sets, respectively. One limitation of this method was the sensitivity of the signal-to-noise ratio during segmentation of the silence and voiced components, where low SNR resulted in false segmentation. In another study Wu et al. (2020) the authors create a model that fits the distribution of genuine speech, i.e., one that takes genuine speech as the input and generates genuine speech as the output. However, if the speech is spoofed, the output will be very different. They propose a genuinization transformer that uses genuine speech features with a convolutional neural network (CNN). The genuinization transformer is then used with an LCNN system for the detection of synthetic speech attacks. The model achieves an EER of 4.07% and a min t-DCF of 0.102% on the ASVspoof2019 dataset using CQCC and LFCC features. However, this method has not been tested across corpora. Following this, the author of Ma et al. (2021) proposes a Conv1D Resblock with a residual connection, which allows the model to learn a better feature representation from raw waveforms. They find that feeding a raw waveform directly into a neural network is adequate. However, the system is only tested against the ASVspoof2019 dataset, and though it achieves an EER of 2.98% against known attacks from that source the presented system's performance was neither reported against PA attacks nor evaluated against cross-corpora.

The work of Tak et al. (2021) is an extension of a previously proposed GAT-ST and uses a raw waveform. This captures discriminative cues in both the spectral and temporal domains. The proposed RawGAT-ST model uses a one-dimensional convolution layer to ingest raw audio. This end-to-end architecture uses feature representation learning and a GAT that learns the relationships between cues at different sub-band and temporal intervals. This model uses the raw waveforms from the ASVspoof 2019 dataset and achieves an EER of 1.06% and a min t-DCF of 0.0335%. However, this model is only tested on one dataset. In another study Liu and Yang (2020) the authors concatenate four sub-features on the ASVspoof2019 dataset. These features, Short-Term Spectral Statistics Information (STSSI), Octave-band Principal Information (OPI), Full-band Principal Information (FPI), and Magnitude-Phase Energy Information (MPEI), are fused to generate the delta acceleration coefficients as features for spoofing detection such as CQSPIC, CQEPIC, and CESPIC. The fusion of the features achieves an EER of 7.63% and a min t-DCF of 0.178%. Although this model's performance is evaluated against both LA and PA attacks it is not evaluated across corpora.

In Yang et al. (2019), the authors propose a heuristic feature extraction method based on Multi Level Transform (MLT), which extracts valuable information from the octave power spectrum for spoofing attack detection. It relies on MLT to extract relevant information from previous DCT results. The authors apply it to the ASVspoof2015 and ASVspoof2017v2 datasets, and achieve an EER of 14.45%. However, it is only tested on two datasets. To fully evaluate the performance of the system it needs to be tested against recent advancements and complex spoofing attacks. This paper, Suthokumar et al. (2019), compares genuine and spoofed speech across different phonemes and shows that specific phonemes (fricatives, nasals, stops, and pauses) are more informative in the detection of replay attacks. It creates four different fusion scoring methods to incorporate phonetic information using phoneme-specific models. This method is tested on the ASVspoof2017 V2 dataset, and achieves an EER of 6.18%. However, it has to be fused with a phenome-independent model for the best results. The work of Wang et al. (2019), which also looks at sound characteristics, creates Voice-Pop, which identifies a live user by detecting the pop noise naturally incurred by a user breathing while speaking close to the microphone. The authors use their own dataset with GFCC features and achieve an EER of 5.4%. However, this method was only tested on one dataset. It needs to be evaluated on multiple datasets,

with different speakers and spoofing attack configurations, and across corpora in order to test the generalizability of the system.

A few solutions use lightweight deep learning systems to detect replay spoofing. A lightweight CNN based on Maximal Feature-Map (MFM) activation, is used to detect replay attacks in Wu et al. (2018). MFM is able to minimize the dimensionality by using the most relevant features for classification. The method described in Wu et al. (2018) is extended in Balamurali et al. (2019) to investigate the efficacy of angular margin-based softmax activation in training a light CNN for cloning and replay spoof detection. It relies on MLT to extract relevant information from previous DCT results. The authors apply it to the ASVspoof2015 and ASVspoof2017v2 datasets, and achieve an EER of 14.45%. However, the generalization of the presented system was not evaluated across corpora.

4.4 Discussion, challenges and the limitations of deep learning counter measures

By analyzing the existing deep learning counter measures it may be observed that the developed solutions have been shown to be effective in detecting synthetic or manipulated speech. However, there are still many challenges and limitations that need to be addressed, such as the need for a large amount of data to train deep learning models. This can be difficult to obtain, especially for under-represented languages or dialects. Additionally, the data used to train models may not be representative of the types of spoofing attacks that are actually encountered in the wild, leading to poor performance in real-world situations. Moreover, one limitation of deep learning-based counter measures is that they can be vulnerable to adversarial attacks. While these may be addressed by using robustness techniques such as adversarial training, this requires even more training data. Overall, while deep learning-based voice spoofing counter measures have shown promise, more research is needed to improve their effectiveness and robustness if they are ever to be deployed in a production system.

4.5 End-to-end solutions

Apart from the progress of handcrafted counter measures, end-to-end solutions to the problem of voice spoofing attacks have been gaining a lot of attention in recent years. These employ advanced signal processing and deep learning techniques to detect the anomalies or inconsistencies associated with voice spoofing attacks. In Dinkel et al. (2017), an end-to-end system was proposed to detect replay attacks against ASVs. This system used a combination of a CNN, LSTM, and DNN to take in raw audio as input and classify the audio into genuine, synthetic/cloned, or replay. The performance of the model was evaluated on the ASVspoof 2015 and BTAS 2016 datasets and achieved a half total error rate (HTER) of 1.56%, compared to the GMM-PLP-39 baseline model at 2.96%. This system may be enhanced in the feature extraction stage by employing better time and frequency filtering networks. Although the proposed system was tested against two datasets, its performance across corpora was not reported. More recently, in Jung et al. (2021), the authors propose a novel heterogeneous stacking graph attention layer that models artifacts spanning heterogeneous temporal and spectral domains with a heterogeneous attention mechanism and a stack node. In concert with a new max graph operation that involves a competitive mechanism and an extended readout scheme, their approach, named ASSIST, achieves an EER of 0.83% and a min t-DCF of 0.0275% on ASVspoof2021. However, the proposed method has only been tested on one dataset. In Tak et al. (2021), a graph attention network

(GAT) is proposed that works by applying a self-attention mechanism to GNNs and modeling graph-structured data. Each node in the graph is weighted according to its relevance to other nodes. GATs can be used to model a specific sub-band or temporal segment using high-level representations extracted from deep residual networks. It achieves a min t-DCF of 0.0089% on the ASVspoof2019 dataset. However, the GAT works on filter bank outputs rather than waveforms, demonstrating the model's dependency on filter bank extraction. Furthermore, it has only been tested on one dataset. One notable work Chen et al. (2023) from ICASSP2023 proposes a graph-based method to enhance spoofing detection in speech systems. The authors introduce the concept of spectro-temporal dependency, capturing the inter-regional relationship between genuine and spoofed speech. Their approach utilizes a graph neural network, incorporating prior knowledge through a graph structure design and edge weighting. An attention mechanism is employed to emphasize critical nodes. The method achieves an impressive equal error rate of 0.58% on the ASVspoof 2019 LA dataset, surpassing competing systems and demonstrating its efficacy in enhancing spoofing detection performance. However, while Chen et al. (2023) demonstrates exceptional performance in detecting logical voice spoofing attacks, it does not report the performance of the proposed solution against other types of voice spoofing attacks, such as replay and adversarial attacks. In another study Xue et al. (2023), a novel self-distillation method is proposed for fake speech detection (FSD), which improves FSD performance without increasing model complexity. The approach involves dividing FSD networks into segments, with the deepest network serving as the teacher model and other networks as student models. A distillation path reduces feature differences between the deepest and shallowest networks. Experimental results on the ASVspoof 2019 LA and PA datasets demonstrate the effectiveness of the proposed method, with significant improvements over the baseline.

In contrast, the authors of Ding et al. (2023) introduce the Speaker Attractor Multi-Center One-Class Learning (SAMO) framework for voice anti-spoofing systems. SAMO addresses unseen attacks by clustering bona fide speech around speaker attractors and separating spoofing attacks in a high-dimensional embedding space. The algorithm co-optimizes bona fide speech clustering and classification during training, enabling anti-spoofing for unenrolled speakers during inference. Experimental results show that SAMO outperforms existing methods, achieving a relative improvement of 38% in EER on the ASVspoof2019 LA evaluation set. However, the paper does not provide information about the performance of the proposed solution on ASVspoof2019 PA attacks, nor does it report the computational complexity resulting from the extraction of high-dimensional embeddings.

The ASVspoof challenge aims to improve voice spoofing detection systems with high accuracy, but overlooks model complexity and latency requirements for real-world implementation. Many top-performing solutions utilize ensemble techniques with deep learning models, which are impractical for resource-restricted voice assistants. To address this limitation, a compact system is proposed in Kwak et al. (2023) which merges skip connections from ResNet and max feature maps from a Light CNN. The optimized single model achieves a replay attack detection EER of 0.30% on the ASVspoof 2019 dataset, outperforming ensemble systems. By incorporating depthwise separable convolutions from MobileNet, the proposed solution significantly reduces the parameter count while maintaining comparable performance. Furthermore, the utilization of Grad-CAM enables the identification of crucial spectrogram regions. This identification contributes to fake data detection.

4.6 Discussion, challenges and the limitations of end-to-end counter measures

Although End-to-end voice spoofing counter measures, the newest approach, aim to detect synthetic or manipulated speech by training a single model to perform both feature extraction and classification, there are significant challenges and limitations that need to be addressed in order for this approach to be practical or effective. End-to-end counter measures require large amounts of data to train, which can be difficult to obtain, itself requiring extensive processing if it is even available. Another limitation is the sensitivity of the model to the quality of the training data; if the data is not representative of the types of spoofing attacks that are actually encountered in the wild, the model may not generalize well to new examples. In addition, end-to-end counter measures are more complex than traditional counter measures, which can make them harder to interpret and difficult to understand. The can also be vulnerable to adversarial attacks, where an attacker manipulates the input speech to evade detection. In addition, explainability and fairness need to be considered when designing end-to-end systems. Explainability factors in transparency and comprehension in the decision-making process, and fairness-oriented systems ensure that the technology is not prejudiced against certain groups or demographics. Implementing explainability and fairness in voice spoofing systems would not only improve their efficacy, it will increase the trust of lay people in such systems. Finally, they require significant computational resources, which can make them difficult to deploy on edge devices with limited computational resources. Overall, while end-to-end voice spoofing counter measures show promise, further research is needed to address these challenges and limitations in order to make them more practical and effective in a real-world deployment.

4.7 Unified solutions

By and large, existing counter measures for voice spoofing attacks exclusively target one type of spoofing attack (e.g., voice replay or SS). In contrast, if ASV systems are ever going to be effective in a real-world scenario, they must by necessity be agnostic to the spoofing type; hence, identifying a single form of spoofing attack is inadequate to ensure the security of ASV systems. Consequently, unified solutions for dealing with all types of spoofing attacks must be developed. Nevertheless, in comparison to single attack detection systems, relatively few unified solutions have been reported to date. One unified solution for voice spoofing detection is presented in Javed et al. (2022). In this work, the authors introduce the ATCoP feature descriptor for the detection of voice presentation attacks, capable of recognizing both LA and PA attacks. Although a good all-around solution, this unified anti-spoofing technique is most effective at detecting single-and multi-order replay attacks. The experimental results show that the presented approach is effective when tested on four distinct datasets, with either replay or clone forgeries. Although this method is evaluated against several datasets, the success of the given ATCoP descriptors is not reported for DF and adversarial attacks. Following this, the authors extend their work and presents a novel hybrid voice spoofing attack, a cloned replay, which may also be used to spoof an ASV system Javed et al. (2021). A cloned replay attack duplicates the audio of a target speaker and transmits it to an ASV system. The author establishes the basis for a spoofing countermeasure capable of detecting multi-order replay, cloning, and cloned-replay attacks through the use of the proposed ATP-GTCC features. This approach is capable of identifying SOTA voice spoofing attacks with a unified solution. However, the presented system

depends entirely on the effective extraction of ATP and GTCC features and has not been tested against DF attacks.

In another study, the author Rostami and Homayounpour et al. Rostami et al. (2021) propose an effective attention branch network (EABN) for detecting LA and PA attacks. The provided technique achieves EERs of 0.86 and 1.89%, and min t-DCFs of 0.02 and 0.50%, against PA and LA attacks, respectively. Despite outperforming SOTA approaches on LA and PA attacks, Rostami et al. (2021) was not tested on deepfake attacks and cross corpora. In contrast, in Lai et al. (2019), SENet and ResNet were combined with statistical pooling to handle anti-spoofing with deeper and faster-trained DNNs. This consisted of a SENet, a median-standard ResNet, a dilated ResNet, and an attentive filtering network. A GNN back-end classifier was implemented using CQCC and LFCC features. It obtained an EER of 0.59% for a PA and 6.7% against LA attacks. Because the data was obtained directly from the ASVspoof2019 dataset with no data augmentation or cross-corpus dataset, the proposed method's performance was not checked against DF attacks and may decline in a real-world scenario. In Chen et al. (2021) the authors presented Emphasized Channel Attention, Propagation, and Aggregation Time-Delayed Neural Networks (ECAPA-TDNN) as their primary model. The authors' intention was to tackle the issue of channel variability by employing an acoustic simulator in order to enhance the original datasets with transmission codecs, compression codecs, and convolutional impulse responses. The presented method attained an EER of 5.46% and a min t-DCF of 0.3094 in the 2021 LA task and an EER of 20.33% on the DF task. Although the presented work prevents several types of attacks, it needs testing and reporting against PA attacks.

4.8 Discussion, challenges and the limitations of unified voice spoofing counter measures

Unified voice anti-spoofing counter measures may well be the wave of the future but there are several limitations that must be considered when using this approach. For instance, they may not generalize well to spoofing attacks that are not represented in the training data. In addition they share a vulnerability to adversarial attacks with end-to-end and handcrafted solutions. Another limitation is a tendency to overfit to the training data due to differentiation boundaries of the spoofing attacks, and as a result they may not generalize well to new examples. Finally, the unified spoofing counter measures have been observed to be biased toward specific types of spoofing attacks. While further research may address these limitations and make them more practical and effective in real-world scenarios, researchers must use caution to avoid unintentional bias.

4.9 Integrated solutions (antispoof & ASV)

The research community for secure voice-enabled systems is currently focused on integrating research efforts on speaker verification and anti-spoofing. Unlike existing anti-spoofing systems, which focus on independently streamlined spoofing detection, speaker verification may also be embedded in order to build a integrated system. New integrated spoofing technologies have been presented that conducted speech spoofing and ASV simultaneously, in addition to unified techniques for cutting-edge spoofing attacks.

The first paper in this regard is Jung et al. (2022), in which the authors propose the spoofing-aware speaker verification (SASV) challenge, which integrates speaker verification and anti-spoofing. In this challenge, the organizers encourage the development of

integrated SASV systems that use new metrics to evaluate joint model performance by releasing official protocols and baseline models. The authors extend speaker verification by including spoofed trials in addition to the standard set of target and imposter trials. Unlike the existing ASVspoof challenge, which focuses on separate spoofing detection and speaker verification systems, SASV aims to develop jointly optimized secure ASV solutions. Open-source, pre-trained spoofing detection and speaker verification models are used in two baseline SASV solutions. Participants have free access to both models and baselines, which can be used to develop back-end fusion approaches or end-to-end solutions. The top performing system reduced the equal error rate of a conventional speaker verification system from 23.83 to 0.13% when tested with target, bonafide non-target, and spoofed non-target trials. The SASV challenge results demonstrate the dependability of today's cutting-edge approaches to spoofing detection and speaker verification. In another work in this domain, Zhang et al. (2022), Zhang et al. develope a probabilistic framework for combining the ASV and countermeasure (CM) subsystem scores. In addition to the probabilistic framework, the authors propose direct inference and fine-tuning strategies, based on the framework, to predict the SASV score. Surprisingly, these strategies reduce the SASV EER of the baseline to 1.53% in the official SASV challenge evaluation trials. The author validates the efficacy of the proposed modules through ablation studies and provides insights through score distribution analysis. However, the proposed system is not tested against DF attacks or across corpora.

The authors of Aljasem et al. (2021) introduce unique sign-modified acoustic local ternary pattern, sm-ALTP, features and an asymmetric bagging-based classifier ensemble with an enhanced attack vector. In addition to a convex function, sm-ALTP is employed to cluster the high- and low-frequency components of the audio frames. The proposed system categorizes several forms of attack on the ASV system, and discriminates between real and spoofed speech samples. Although the presented method is proven against several spoofing attacks and across corpora, more testing of the system's efficacy against complex spoofing attacks using compression and different encoding artifacts is required.

The author of Liu et al. (2022), expresses concern about the improvement of spoofing robustness in automatic speaker verification (ASV) systems in the absence of a separate countermeasure module. To address this issue, the ASVspoof 2019 baseline model is used, in accordance with the back-end machine learning classifier's probabilistic linear discriminant analysis (PLDA). Three unsupervised domain adaptation techniques are used to optimize the back-end using audio data from the ASVspoof 2019 dataset training partition. The results show significant improvement in both the logical and physical access scenarios, particularly in the latter, where the system is attacked by replayed audio, with maximum relative improvements of 36.1% and 5.3% in bonafide and spoofed cases, respectively. However, absolute error rates on spoof trials remain too high. This demonstrates the challenge of making a conventional speaker embedding extractor with a PLDA back-end work on a mix of bonafide and spoofed data.

Over the last few years, research has improved the performance of ASVs and countermeasure systems, resulting in low EERs for each system. However, research on the joint optimization of both systems is still in its early stages. This paper, Teng et al. (2022), proposes a Spoof-Aggregated-SASV (SA-SASV), an ensemble-free, end-to-end solution for developing an SASV system with multi-task classifiers. The SA-SASV system is further optimized by multiple losses and more flexible training set requirements, and is trained using the ASVspoof2019-LA dataset. SA-SASV EER results show that training in complete ASV and countermeasure datasets can improve model performance even further. The results show that the SA-SASV feature space outperforms previously published approaches

in terms of distinguishing spoof attacks and speaker identification. Furthermore, the SASASV EER is reduced from 6.05%, produced by previous SOTA approaches, to 4.86% when no ensemble strategy is used. Although the article argues that a larger set of data and distinct encoders will further improve the EER, the proposed solution was not tested across corpora.

4.10 Discussion, challenges and the limitations of integrated voice spoofing counter measures

There are some significant downsides to this strategy that must be considered. They suffer from issues that plague more general anti-spoofing solutions, i.e., a considerable quantity of data is required, a susceptibility to adversarial attacks, and in a real-world scenario, when speech may be influenced by background noise, reverberation, and other distortions, integrated counter measures may not work effectively. In addition, they can be computationally intensive, making them difficult to deploy on edge or other devices with limited processing capabilities. Coupled with this is the need to achieve a balance between ASV performance and anti-spoofing detection. Lastly, they need a multi-task learning strategy, which may be difficult to design and optimize. Overall, while integrated speech anti-spoofing remedies offer promise, a tremendous amount of research is needed to solve these limitations and difficulties in order to make them practical and successful.

4.11 Adversarial attack counter measures

Adversarial attacks use small perturbations in audio samples, such as the addition of Gaussian noise, to pose a threat to ASV anti-spoofing models. Though options are only beginning to be explored in the literature, a few methods stand out. One, described in Wu et al. (2020a), uses self-supervised learning to leverage knowledge of unlabeled data, which significantly improves performance. This model creates high-level representations, extracted by the self-supervised model, and a layer-wise noise to signal ratio (LNSR) in order to quantify and measure the effectiveness of deep models in countering adversarial noise. During training, different masking methods are applied to 15% of the audio samples to create a model that reconstructs audio. This defense, called Mockingjay, is passive, i.e., it does not proactively seek out adversarial attacks, and is only tested on the Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM) attacks. Also, as epsilon (the strength of an attack) increases, this defense starts to weaken and eventually fails. Both white-box and black-box scenarios were tested. Another defensive method, Wu et al. (2020b), uses spatial smoothing, filtering, and adversarial training. This method has only been tested against PGD attacks and may perform poorly under other attacks due to changing patterns on the sample. These defense methods improved the robustness of spoofing counter measures against PGD attacks, but they are still limited in their overall defensive capability because adversarial training depends on familiar attacks.

Turning to the description of new adversarial attacks, Xie et al. Xie et al. (2021) describe a quick and universal adversarial attack on three audio processing systems: speech command recognition, speaker recognition, and environmental sound classification. The proposed approach, which uses Wave-U-Net and class-wise feature embedding maps, can launch a fast audio adversarial attack targeting any speech command using a

unified generative model within a single pass of feed-forward propagation. This results in an adversarial perturbation generation speedup of up to 214 times faster than SOTA solutions. Furthermore, because it is based on a fast audio adversarial perturbation generator (FAPG), the proposed universal audio adversarial perturbation generator (UAPG) produces universal adversarial perturbations that can be applied to any benign audio input. Extensive trials show the efficacy of the suggested FAPG and UAPG. Finally, the proposed UAPG provides universal adversarial perturbations that outperform SOTA solutions in terms of attack performance. In another paper on adversarial attacks, Li et al. (2020), the authors examine the vulnerability of speaker recognition systems by introducing practical and systematic adversarial attacks. The efficiency of the system built on X-vectors and cutting-edge deep neural networks is specifically examined in this study. In particular, Liu et al. add well-crafted, inconspicuous noise to the original audio when building the adversarial attack. They claim that by injecting inconspicuous noise, the speaker identification system may be induced to make incorrect predictions, and may even cause the audio to be perceived as any adversary-desired speaker. Furthermore, the attack incorporates estimated room impulse response (RIR) into the adversarial example training process, resulting in realistic adversarial examples that may be broadcast over the air and remain effective. An extensive experiment using a public dataset of 109 speakers demonstrates the effectiveness of the proposed attack, which achieves high accuracy for both digital (93%) and practical over-the-air (50%) attacks.

4.12 Discussion, challenges and the limitations of integrated voice spoofing counter measures

Adversarial attacks on voice spoofing are a significant concern for the security of voice-based systems, however, there are challenges in detecting and preventing these attacks. One of the main limitations is the lack of accurate ground truth data for the systems to learn from, which makes it difficult to train robust classifiers. Additionally, the adversarial examples generated through such attacks are often indistinguishable from real speech, making it challenging to detect them in real-time. These limitations emphasize the importance of ongoing research and development in the area of voice spoofing and adversarial attacks to improve the security of voice-based systems.

4.13 Other voice spoofing solutions

In addition to these solutions, transfer learning, generative adversarial networks (GANs) and auto-regressive networks have recently made significant contributions to audio spoofing detection Cai et al. (2018); Saito et al. (2018); Kwon et al. (2023); Ba et al. (2023). These approaches have not been extensively explored but have shown promising results in small-scale experiments. For instance, in one study Cai et al. (2018), GANs were used to synthesize spoofing attacks on speaker recognition systems, demonstrating the effectiveness of GAN-generated adversarial examples. Another study Saito et al. (2018) proposed a method incorporating GANs to improve the quality of generated speech parameters. The results showed significant improvements in speech quality compared to conventional training algorithms. Furthermore, a study on audio adversarial examples Kwon et al. (2023), applied audio style transfer learning to defend against attacks, achieving improved accuracy against adversarial examples. Finally, a study on cross-language deepfake detection Ba et al. (2023) explored transfer learning techniques to detect fake audio across different

Table 4 A summary of databases in terms of training, development and evaluation speech samples

AVspoof						
Track	Genuine		Spoof LA Attacks		Spoof PA Attacks	
Training	4973		38580		17890	
Development	4995		38580		17890	
Evaluation	5576		43320		20060	
ReMASC						
Environment	Speaker		Genuine		# Replayed	
Outdoor	12		960		6900	
Indoor 1	23		2760		23104	
Indoor 2	10		1600		7824	
Vehicle	10		3920		7644	
Total	55		9240		45472	
VSDC						
Track	Environment		Sample Rate		# Utterances	
Bonafide	Office Desk,				4000	
Replay-1PR	Living Room,		96K		4000	
Replay-2PR	Kitchen Table				4000	
Total	25		–		12000	
ASVspoof2015						
	Speaker		Genuine		Spoofed	
Track	Male	Female	# Samples		# Samples	
Training	10	15	3750		12625	
Development	15	20	3497		49875	
Evaluation	20	26	9404		193406	
ASVspoof2017						
Track	Speaker		Genuine		Spoofed	
Training	10		1507		1507	
Development	8		760		950	
Evaluation	24		1298		12008	
ASVspoof2019						
	Speaker		LA Attacks		PA Attacks	
Track	Male	Female	Genuine	Spoofed	Genuine	Spoofed
Training	8	12	2580	22800	5400	48600
Development	8	12	2548	22296	5400	24300
Evaluation	-	-	71747		137457	

language domains, demonstrating the adaptability of models to unseen algorithms. Overall, these research works highlight the potential of transfer learning and GANs in enhancing audio spoofing detection and addressing related challenges.

Table 5 A summary of the RedDots, VoxCeleb and BioCPqD-PA databases in terms of languages, No of speakers and utterances

Dataset	Language	Attacks	Speakers	Utterances
RedDots	5	Replay	3750	5000
VoxCeleb (V1)	6+	Mimicry	1251	100,000
VoxCeleb (V2)	6+	Mimicry	7000+	1,000,000
BioCPqD-PA	1	Replay	222	418,940

5 Publicly available voice spoof detection datasets

This section of the paper discusses and identifies the various datasets used in cutting-edge ASV systems. Early stages of voice spoofing detection research involved speech and speaker recognition databases, i.e., YOHO Kreuk et al. (2018), NIST Alegre et al. (2013), and WSJ Ergünay et al. (2015). However, to accurately account for research progress, it was evident that there was a dire need for a common dataset as well as a performance metric to evaluate spoofing counter measures, which was discussed in depth and addressed at the INTERSPEECH 2013 special session on spoofing and ASV counter measures. This motivated the research community to organize the first Automatic Speaker Verification Spoofing and counter measures Challenge, ASVspoof, in 2015, which took place at INTERSPEECH that year. The dataset released for this initial challenge included two types of spoofing attacks: SS and VC. In the years following, three additional ASVspoof challenges were organized: ASVspoof 2017, ASVspoof 2019, and ASVspoof 2021, each with publicly available datasets for download.

There are several common publicly available datasets which are used by voice PAD researchers. This section will briefly cover the existing datasets (2015–2022) used for spoof detection and countermeasure development. Details of the publicly available datasets which address spoofing attacks are presented in Tables 4 and 5.

5.1 Datasets used in this experimental analysis

The **ASVspoof 2019** challenge Wang et al. (2020) is an extension of the previously held ASVspoof challenges. It focuses on counter measures for all three major attack types, namely, SS, VC, and replay. It is divided into the logical access (LA) and physical access (PA) subsets. LA contains TTS and VC spoof speech samples, and PA has replay-spoof speech. Both of these subsets are further partitioned into training, development, and evaluation subsets. The training subset is generated by 8 males and 12 females, the development subset by 4 males and 6 females, and the evaluation subset by 21 males and 27 females. The spoofing speech signals are generated using one of the two VC and four SS algorithms. The ASVspoof datasets are the de facto standard for the detection of spoofing attacks and are thus a natural inclusion in this study. **ASVspoof 2021** Delgado et al. (2021) is the fourth offering in the series of spoofing challenges. ASVspoof 2021, in particular, is divided into three sub-challenges. The first is a logical access sub-challenge that builds on the 2019 challenge by emphasising robustness to channel variation; the second is a physical access sub-challenge, similar to the 2019 setup, but with recordings made in real-world physical environments. The third is a speech deepfake detection sub-challenge (no ASV). ASVspoof 2021 contains technically difficult data to encourage broad generalization of counter measures. The logical access (LA) task includes the computation and transmission

of text-to-speech (TTS) and VC attacks. In comparison to ASVspoof 2017, the physical access (PA) task includes genuine and replayed samples but with a more tightly controlled setup. The new speech deepfake task (DF), similar to the LA task, includes compressed audio. The protocols used in the training and development sections are the same as in ASVspoof 2019, and therefore ASVSpoof 2021 does not include development or training subsets. In particular, the logical access and physical access subsets of ASVspoof2021 include 181, 566 and 943, 110 speech samples, respectively, and the speech deepfake set includes 611, 829 speech samples. ASVspoof2021 also introduces new metrics for the evaluation partition, including a slightly revised t-DCF metric for the LA and PA tasks. However, EER is still used for the evaluation of the DF task. Like ASVspoof2019, this dataset is a community standard. In addition, the DF task, included for the first time in ASVspoof2021, increases the breadth of the detection challenges, and was a good fit for this study's focus on generalizability.

The **Voice Spoofing Detection Corpus (VSDC)** dataset Baumann et al. (2021) comprises first-order (once replayed) and second-order (twice replayed) samples over real audio recordings. In terms of environments, record and playback devices, speakers, setups, and replay scenarios, the VSDC dataset represents a diversified replay spoofing detection corpus. The VSDC dataset, in particular, involves the utilization of 35 microphones, 25 distinct recording setups, and 60 various playback devices for first- and second-order replays in order to obtain a total of 12,000 samples from 19 speakers. This dataset was generated with the intention of simulating these attacks in a controlled environment. While the primary goal of VSDC is to detect multi-hop replay attacks, it may be used to simulate a variety of PA scenarios, e.g., typical replay attacks, the influence of different microphones and surroundings on an audio file, or how an individual's vocal range affects the accuracy of a voice control system. The details of tracks, environment and number of speech samples for VSDC is presented in Table 4, and details of the development architecture can be found in Baumann et al. (2021). VSDC is included in these experiments because it fills a gap in the otherwise homogeneous ASVspoof testing. To adequately test the generalizability of the counter measures it is necessary to present a variety of scenarios, and VSDC accomplishes this by introducing the multi-order replay attack not found in ASVspoof.

In the experimental setup, we selected voice spoofing datasets characterized by an imbalanced distribution of samples and utilized them in their original form. This deliberate choice is made to promote a fair comparison to real-world spoofing scenarios, where attack detection models often encounter imbalanced datasets. By incorporating these imbalanced datasets, we were able to evaluate the performance of SOTA counter measures under conditions that closely resembled practical voice spoofing applications. This approach enabled us to thoroughly assess the robustness and efficacy of the SOTA techniques in effectively handling an imbalanced data distribution, which is prevalent in the voice spoofing detection domain.

To ensure the generalizability of the evaluated counter measures, we conducted evaluations across corpora. This comprehensive evaluation strategy enhances the reliability and applicability of our findings. By employing imbalanced datasets and conducting rigorous evaluation our experimental methodology aims to provide valuable insight into the real-world performance of the SOTA counter measures specifically tailored for voice spoofing detection. These insights contribute to the advancement of practical solutions and their effective deployment in voice spoofing detection applications.

5.2 Other publicly available datasets and ASV challenges

The **Spoofing and Antispoofing (SAS)** corpus Wu et al. (2015a) contains a wide range of spoofed speech samples generated using nine different approaches, two of which are SS, and the other seven are VC. The database contains two protocols: one for testing the ASV system and another for generating spoofed speech sounds. Periods of silence not found in natural speech were removed from the samples, resulting in a more realistic SS and VC spoof corpus. The SAS corpus contains speech produced using various spoofing methods, represented in 300,000 samples of each type.

The **RedDots** project Lee et al. (2015) was launched as a follow-up to a special session at INTERSPEECH 2014, which was designed to bring together research efforts to investigate potential approaches and gain a better understanding of speaker/channel phonetic variability. The goal of the RedDots project was to collect speech data through mobile crowd sourcing, which allows for a larger population and greater diversity. At the time of our investigation, the project had 89 speakers from 21 countries, 72 men and 17 women, for a total of 875 complete sessions. Each session is limited to two minutes, with a total of 24 sentences (10 common, 10 unique, 2 free choices, and 2 free texts) for each session. Following that, the replayed RedDots database Kinnunen et al. (2017b) was created by re-recording the original corpus utterances under different environmental conditions.

The **AVspoof** dataset, not to be confused with the ASVspoof series of datasets, is designed to assist ASV systems in the development of anti-spoofing techniques, and was first used in the BTAS 2016 Korshunov et al. (2016); Ergünay et al. (2015) challenge. It includes replay spoofing attacks in addition to synthetic speech and VC spoofing attacks. The replay attacks are generated by various recording devices. The SS attacks are generated using the Hidden Markov Model (HMM) and Festvox methods, which account for the vast majority of the VC attacks. The sessions are recorded by participants in a variety of environments and with a variety of recording devices. Speakers are instructed to read out sentences, phrases, and speak freely about any topic for 3 to 10 min. To make competition more difficult, "unknown" attacks are included in the test set Korshunov et al. (2016).

VoxCeleb Nagrani et al. (2017) is an audio-visual dataset comprised of short clips of human voices extracted from YouTube interview videos. Each segment lasts at least 3 s. VoxCeleb features speech from people of various ethnicities, accents, professional backgrounds, and ages. 61% of speakers are male and 39% are female. The data was collected randomly, with ambient noise, laughter, overlapping speech, pose deviation, and a variety of lighting conditions. This dataset is available in two versions: VoxCeleb1 Nagrani et al. (2017) and VoxCeleb2 Chung et al. (2018). Each has audio files, face clips, metadata about speakers, and so on, in the training and testing sets.

The **voicePA** dataset was created with the assistance of the AVspoof dataset. Its bonafide data is a subset of the AVspoof dataset's genuine data, uttered by 44 speakers in four recording sessions held in various settings Wang et al. (2020). These sessions were recorded using high-quality microphones from a laptop, a Samsung S3, and an Apple smartphone 3 GS. Spoofed data consists of 24 different types of presentation attacks that are captured using five devices in three different environments. These spoof utterances are based on real-world data.

The **Portuguese language BioCPqD-PA** dataset Korshunov et al. (2018) was collected by recording 222 people in a variety of environmental conditions, and is comprised of 27,253 authentic recordings and 391,687 samples which have been subjected to

a presentation attack. One laptop was used with 24 different setups consisting of 8 loudspeakers and 3 microphones, while another single laptop was used to capture real data.

The **ASVspoof 2015** Challenge database was the first significant release for research into spoofing and counter measures Wu et al. (2015b). For LA attacks, the database contains natural and spoofed speech generated by SS and VC. The evaluation subset is made up of both known and unknown attacks. The known attacks contain the very same five algorithms that were used to generate the development dataset and are thus referred to as known (S1-S5) attacks. Other spoofing algorithms are included in unknown (S6-S10) attacks that are directly used in the test data.

The **ASVspoof 2017** Challenge dataset Kinnunen et al. (2017c) was constructed on the RedDots dataset Lee et al. (2015). This dataset Kinnunen et al. (2017c) contains replayed data samples with text-dependent speech, and includes the voices of 42 different speakers, recorded using 61 combinations of distinct recording devices, replay devices, and environmental conditions. It was collected over the course of 179 sessions. The original ASVspoof 2017 database Kinnunen et al. (2017c) contained some inconsistencies, but the issue was resolved in ASVspoof 2017 Version 2.0 Delgado et al. (2018). In addition to the corrected data, a more detailed description of recording and playback devices, as well as acoustic environments, was provided Delgado et al. (2018).

The **Realistic Replay Attack Microphone Array Speech Corpus (ReMASC)** is a database of speech recordings developed to support research into and the security of voice-controlled systems. It is comprised of authentic and replayed recordings of speech samples, captured in actual circumstances, and utilizes cutting-edge voice assistant development kits. In particular, it contains recordings from four systems, each with a different transmitter and receiver, under a range of atmospheric situations, with varying levels of background noise and relative speaker-to-device locations. This is the first database that was specifically developed to safeguard voice-controlled systems (VCS) from various types of replay attacks in varied contexts.

6 Performance evaluation parameters

This subsection provides an overview of the evaluation criteria used in this survey, with a specific focus on Equal Error Rate (EER) and the minimum Tandem Detection Cost Function (min t-DCF). These metrics provide a robust assessment of system performance and facilitate ASV-centric evaluation. By incorporating EER and min t-DCF as the primary evaluation metrics in our experiments, we ensure consistency with previous studies and enable meaningful comparisons of the proposed counter measures. A detailed discussion of EER and min t-DCF is presented below:

6.1 Equal error rate (EER)

The Equal Error Rate (EER) is used to evaluate the performance of Automatic Speaker Verification (ASV) systems. An ASV system categorizes identities as either approved or rejected, and there are four possible outcomes: True Acceptance (TA), True Rejection (TR), False Acceptance (FA), and False Rejection (FR). TA and TR are desirable outcomes, while FA and FR are not. These outcomes are determined based on a preset threshold Todisco et al. (2017). EER is the threshold value at which the False Acceptance Rate

(FAR) and False Rejection Rate (FRR) are equal. It provides a comprehensive assessment of the ASV system's performance, taking into account both types of errors (FA and FR).

6.2 Tandem detection cost function (t-DCF)

The ASVspoof challenge series was established to drive research in anti-spoofing for automated speaker verification (ASV) systems. Previous challenge editions used the equal error rate (EER) metric to evaluate spoofing counter measures (CMs) separately from ASV, but this approach had drawbacks. To address these issues, the community has transitioned from a CM-centric to an ASV-centric evaluation approach using a new measure called the tandem detection cost function (t-DCF), which consists of six components divided into two parts: false alarm and miss costs, and the prior probability of the target and spoof trials. The results of t-DCF, as presented in the study by Kinnunen et al. (2018), support the inclusion of this DCF-based metric in the future road-map of ASVspoof challenges and other bio metric anti-spoofing evaluations.

7 Experimental configuration and prerequisites

In this section, we report the results of the experiments performed to determine the effectiveness of the SOTA counter measures. The effectiveness of each countermeasure was evaluated by testing it on the training and evaluation sets of the chosen datasets before performing cross-corpus evaluation, where the method is tested by training on one dataset and testing on another. For a fair comparison of the counter measures, we test their effectiveness against four different ML and DL models and two diverse datasets, VSDC and ASVspoof2019. While other datasets were considered, ASVspoof was chosen for its use as the industry standard, and VSDC provides a suitable contrasting configuration suitable for evaluating the generalizability of the tested counter measures. Although some non-English language datasets are available, they were ultimately rejected due to age or lack of standardization or spoof complexity. Simulation on non-English datasets remains an interesting topic and will be reserved for future work if a suitable dataset can be found.

7.1 Models used in the experimental analysis

The experimental analysis in this work was conducted with four distinct classifiers: two machine learning approaches and two deep learning approaches. The classifiers employed were a Gaussian Mixture Model (GMM) Todisco et al. (2019), a Support Vector Machine with Radial Basis Function (SVM-RBF) Aljasem et al. (2021), a Convolutional Neural Network (CNN) Jung et al. (2019), and a CNN-Gated Recurrent Unit (CNN-GRU) Jung et al. (2019).The selection of these classifiers was determined by several criteria. Our first priority was to utilize the conventional baseline classifiers from the voice spoofing community, i.e., the GMM and SVM, and deep learning, i.e., the CNN and CNN-GRU. Secondly, we sought to encompass a diverse range of potential avenues for comparison. Lastly, the availability of source code for these models was also taken into consideration.

Additionally, it is worth noting that we expanded the baseline GMM model, as provided by ASVspoof2019, to evaluate the performance of all of the handcrafted front-end features for CQCC and LFCC. It was trained for ten iterations, and the score of the speech samples

was computed using a log-likelihood ratio. EER and min t-DCF were used to assess the effectiveness of each countermeasure in alignment with the current ASVspoof challenges. Next, an SVM classifier with an RBF kernel was used to evaluate the performance of the SOTA counter measures. The GMM and SVM classifiers were chosen because they had the best reported results for the respective datasets. For example, the SVM-RBF classifier Aljasem et al. (2021) achieved the best performance on the VSDC dataset, whereas CQCC-GMM and LFCC-GMM were the baseline classifiers for the ASVspoof2019 competition. In addition, the CNN and CNN-GRU Jung et al. (2019) models were used to determine the performance of the SOTA counter measures on deep learning-based classifiers. The effectiveness of each countermeasure is described below in the experimental analysis section VII. When testing handcrafted features, the two-dimensional features were converted to one-dimensional through the use of the mean average of the retrieved features.

7.2 Experimental setup and hardware requirements

The setup for the experiments consisted of a pipeline of feature extraction and then evaluation of the extracted features on the ML and DL-based classifiers. The feature extraction and classifiers were run on Oakland University's Matilda High Performance Cluster (HPC). The standard compute nodes were used for feature extraction and model training, each of which consists of 192 GB of RAM and 40 CPU cores at 2.50 GHz. The models (GMM, SVM, CNN, and CNN-GRU) were trained and tested using the HPC's GPU nodes, made up of four NVIDIA Tesla V100 16 G GPUs, 192 GB of RAM, and 48 CPU cores running at 2.10 GHz. The VLFEAT@matlab API was used for the GMMs, which consisted of 512 Gaussian clusters.

8 Experimental analysis of counter measures

This section presents the experimental results of the counter measures against four different classifiers (GMM, SVM, CNN, and CNN-GRU). The counter measures are trained and evaluated using the ASVspoof2019 and VSDC datasets during the experiment, and the results are reported accordingly in Tables 6 to 14. The bold figures present in Tables 6, 7, 8, 9, 10, 11, 12, 13, 14 hold notable significance, as they serve as markers for the lowest and optimal results obtained by the countermeasures under evaluation.

8.1 SOTA features with a GMM classifier

In this experiment, we evaluate the performance of the SOTA countermeasure with the baseline GMM classifier provided by the ASVspoof Wu et al. (2017) challenge community. The baseline GMM with LFCC and CQCC features was expanded to encompass the SOTA handcrafted features and the results of this experiment is presented in Table 6.

8.1.1 Results on the ASVspoof2019 dataset

The counter measures were considered and analyzed for LA and PA attacks against the ASVspoof19 dataset development and evaluation speech samples, and the results are reported in Table 6. They demonstrate that each countermeasure achieved a better EER and

Table 6 Experimental performance of the counter measures with ASVspoof2019 and VSDC corpus on GMM based classifier

Feature	VSDC	ASVspoof2019	Development		Evaluation	
	EER	Attack access	EER	t-DCF	EER	t-DCF
LFCC Zhou et al. (2011)	36.33	ASVspoof-LA	2.70	0.06	8.08	0.21
		ASVspoof-PA	11.9	0.25	13.5	0.30
CQCC Todisco et al. (2017)	20.0	ASVspoof-LA	0.43	0.02	9.57	0.23
		ASVspoof-PA	**4.87**	0.19	11.0	0.24
LPCC Wong and Sridharan (2001)	5.86	ASVspoof-LA	2.31	0.06	10.17	0.28
		ASVspoof-PA	38.5	0.83	46.6	0.98
MSRCC Tapkir et al. (2018)	54.73	ASVspoof-LA	9.45	0.19	10.9	0.28
		ASVspoof-PA	12.7	0.28	15.7	0.38
PSRCC Tapkir et al. (2018)	55.11	ASVspoof-LA	9.13	0.14	10.7	0.21
		ASVspoof-PA	12.7	0.28	15.7	0.38
SCFC Kua et al. (2010)	45.49	ASVspoof-LA	14.7	0.39	20.6	0.54
		ASVspoof-PA	17.6	0.35	21.6	0.47
SCMC Kua et al. (2010)	49.91	ASVspoof-LA	**0.01**	0.54	5.91	0.15
		ASVspoof-PA	12.5	0.27	13.9	0.33
MFCC Muda et al. (2010)	49.90	ASVspoof-LA	7.06	0.16	10.56	0.25
		ASVspoof-PA	11.5	0.23	13.7	0.32
IMFCC Chakroborty and Saha (2009)	50.00	ASVspoof-LA	0.04	**0.01**	10.9	0.24
		ASVspoof-PA	12.8	0.29	13.7	0.32
RFCC Saratxaga et al. (2012)	48.87	ASVspoof-LA	2.71	0.07	8.06	0.22
		ASVspoof-PA	11.8	0.26	13.9	0.33
RPS Saratxaga et al. (2012)	55.21	ASVspoof-LA	9.28	0.19	11.9	0.29
		ASVspoof-PA	15.3	0.32	14.0	0.36
SSFC Tapkir et al. (2018)	53.11	ASVspoof-LA	8.10	0.16	10.3	0.27
		ASVspoof-PA	13.8	0.29	13.9	0.31
GTCC Valero and Alias (2012)	56.00	ASVspoof-LA	9.25	0.17	10.8	0.24
		ASVspoof-PA	12.7	0.28	15.7	0.38
APGDF Rajan et al. (2013)	40.29	ASVspoof-LA	0.22	**0.01**	5.75	**0.14**
		ASVspoof-PA	8.66	**0.17**	**10.6**	0.25

min t-DCF against the development set of each spoofing attack (LA and PA), but performance of the each countermeasure declined when tested against the evaluation set. In particular, when evaluated on LA attacks in the ASVspoof2019 development set, the SCMC Kua et al. (2010) countermeasure achieved the lowest EER, 0.01%, while the lowest min t-DCF, also 0.01%, was obtained by the IMFCC Chakroborty and Saha (2009) and APGDF Rajan et al. (2013) counter measures. However, in evaluation set testing, the APGDF Rajan et al. (2013) countermeasure achieved a lower EER and m-t-DCF, 5.75 and 0.14%, respectively, in contrast to IMFCC's 10.9 and 0.24%.

When considered across the development and evaluation subsets, the results demonstrate that the APGDF Rajan et al. (2013) and CQCC Todisco et al. (2017) counter measures outperform all other features with an overall lower EER and min t-DCF. In addition, some counter measures perform optimally against the LA attacks but failed to perform optimally when replay artifacts existed in the speech sample. For instance, the SCMC Kua

et al. (2010) counter measures obtained the best EER of 0.01% and the second best EER of 5.91% for the development and evaluation test sets, however, SCMC performed poorly when faced with PA attacks. Similarly, the EER of APGDF Rajan et al. (2013) countermeasure was higher than SCMC in the development set but achieved the overall best results in the rest of the testing.

8.1.2 Results on the VSDC dataset

The performance of these SOTA counter measures was also tested with the VSDC dataset, and the results are shown in Table 6. The results show that the LPCC Wong and Sridharan (2001) countermeasure exhibits the best, and CQCC Todisco et al. (2017) obtains the second best EER, 5.86 and 20.0%, respectively. The GTCC countermeasure Valero and Alias (2012) obtained the highest EER of 55.0%, which demonstrates the inadequacy of this countermeasure in the presence of speech samples with distinct microphonic discrepancies. Except for LFCC Zhou et al. (2011), none of the handcrafted features achieved an EER lower than 40.0%, rendering them effectively unsuitable as anti-spoofing counter measures.

8.1.3 Discussion of the GMM experiments

The experimental results showed that each countermeasure performed significantly differently against SS (SS), VC, and replay attacks, and demonstrated that the samples with replay attacks were much more difficult to detect in comparison with the SS and VC samples. During evaluation it was observed that the performance of the SOTA counter measures declined drastically in the presence of replay artifacts in the speech samples.

From these results, we may infer that countermeasure effectiveness is totally dependent on dataset characteristics. In the ASVspoof2019 dataset, for instance, while APGDF Rajan et al. (2013) features produced the best results, on the VSDC dataset its effectiveness decreased by nearly an order of magnitude. Similarly, LPCC Wong and Sridharan (2001) did not generate effective results for the ASVspoof datasets, particularly the ASVspoof-PA assessment set, but produced the best results in VSDC testing. Furthermore, we can see that the front-end features that work well against LA attacks did not perform well against PA attacks. However, the APGDF Rajan et al. (2013), CQCC Todisco et al. (2017), and LPCC Wong and Sridharan (2001) features produced comparable results independent of the dataset's specific configuration.

8.2 SOTA features with an SVM classifier

In this experiment, we evaluate the effectiveness of SOTA counter measures using an SVM classifier. The SVM classifier was shown in our prior work to be superior to the GMM classifier Aljasem et al. (2021). An SVM with an RBF distribution was utilized in these experiments. The RBF distribution was selected after a thorough analysis of all feasible distributions. The scikit-learn Pedregosa et al. (2011) machine learning library was used to train and test the SVM classifier. In addition, a classification report Pedregosa et al. (2011) was used to compute the precision, recall, F1-score, and accuracy of the counter measures.

Table 7 Experimental performance of the counter measures with VSDC dataset and SVM based classifier

Feature	EER	Precision	Recall	F1-score	Accuracy
LFCC Zhou et al. (2011)	0.49	0.88	0.74	0.79	0.88
CQCC Todisco et al. (2017)	**0.27**	**0.92**	**0.85**	**0.88**	**0.93**
LPCC Wong and Sridharan (2001)	0.44	0.88	0.77	0.81	0.89
PSRCC Tapkir et al. (2018)	0.40	0.79	0.76	0.77	0.86
MSRCC Tapkir et al. (2018)	0.46	0.84	0.76	0.79	0.86
SCFC Kua et al. (2010)	0.80	0.39	0.50	0.44	0.78
SCMC Kua et al. (2010)	0.53	0.87	0.72	0.76	0.87
MFCC Muda et al. (2010)	0.46	0.82	0.75	0.77	0.86
IMFCC Chakroborty and Saha (2009)	0.71	0.88	0.64	0.67	0.84
RPS Saratxaga et al. (2012)	0.40	0.76	0.78	0.77	0.85
RFCC Saratxaga et al. (2012)	0.51	0.87	0.73	0.76	0.87
GTCC Valero and Alias (2012)	0.42	0.82	0.76	0.78	0.86
APGDF Rajan et al. (2013)	0.43	0.89	0.77	0.81	0.89
SSFC Tapkir et al. (2018)	0.44	0.87	0.77	0.80	0.88

8.2.1 Results on the ASVspoof2019 dataset

In the case of the ASVspoof2019 dataset, each of the SOTA counter measures was tested against the SVM classifier using the development and evaluation subsets, and the results are reported in Table 8. These demonstrate that in the testing of LA attacks that exist in the development set and evaluation set, the CQCC Todisco et al. (2017) and IMFCC Chakroborty and Saha (2009) counter measures achieved comparable performance, with EERs of 0.30 and 0.29%, respectively. Similarly, for the evaluation set, the CQCC Todisco et al. (2017) and IMFCC Chakroborty and Saha (2009) counter measures achieved the best EER of 0.69 and 0.76%, respectively. Although these EER scores are not good enough for any spoofing system, none of the other counter measures achieved a lower EER.

8.2.2 Results on the VSDC dataset

The results of this experiment on the VSDC dataset are shown in Table 7, and demonstrate that the CQCC Todisco et al. (2017) feature proved to have the most robust front-end attributes of the SOTA counter measures. CQCC features, in particular, achieved the lowest EER of 0.27%, the highest precision, recall, and F1-scores of 0.92, 0.85, and 0.88%, respectively, as well as the highest accuracy of 0.93%. Spectral features such as SCFC Kua et al. (2010), on the other hand, performed poorly in the case of the VSDC dataset, which contains audio samples with micro-phonic inconsistencies. SCFC had the lowest precision (0.39%), recall (0.50%), and F1-score (0.44%), and achieved an accuracy of 0.78%. Similarly, the SCMC Kua et al. (2010) features were found to be the second worst classifier with a 0.53% EER. This indicates that the spectral centroid magnitude and frequency-based coefficients failed to perform effectively when microphonic distinction (as in the case of VSDC) exists in the speech sample. In contrast, the experiments demonstrate that constant Q-cepstral-based coefficient successfully overcame the microphone fluctuations.

Table 8 Experimental performance of the counter measures with ASVspoof2019 and an SVM based classifier

Feature	Dataset	Development		Evaluation	
		EER	Acc	EER	Acc
LFCC Zhou et al. (2011)	ASVspoof19-LA	0.58	0.90	0.91	0.12
	ASVspoof19-PA	0.72	**0.75**	0.83	0.11
CQCC Todisco et al. (2017)	ASVspoof19-LA	0.30	**0.95**	**0.69**	**0.29**
	ASVspoof19-PA	**0.42**	0.29	**0.80**	**0.19**
LPCC Wong and Sridharan (2001)	ASVspoof19-LA	0.50	0.88	0.80	0.15
	ASVspoof19-PA	0.61	0.20	0.81	0.19
SCFC Kua et al. (2010)	ASVspoof19-LA	0.50	0.89	0.90	0.16
	ASVspoof19-PA	0.75	0.19	0.89	0.14
SCMC Kua et al. (2010)	ASVspoof19-LA	0.49	0.93	0.88	0.12
	ASVspoof19-PA	0.75	0.20	0.86	0.15
PSRCC Tapkir et al. (2018)	ASVspoof19-LA	0.49	0.89	0.89	0.11
	ASVspoof19-PA	0.84	0.25	0.89	0.12
MSRCC Tapkir et al. (2018)	ASVspoof19-LA	0.51	0.90	0.91	0.11
	ASVspoof19-PA	0.84	0.25	0.89	0.12
MFCC Muda et al. (2010)	ASVspoof19-LA	0.50	0.89	0.80	0.17
	ASVspoof19-PA	0.76	0.22	0.86	0.16
IMFCC Chakroborty and Saha (2009)	ASVspoof19-LA	**0.29**	0.94	0.76	0.24
	ASVspoof19-PA	0.51	0.26	0.82	0.18
RFCC Saratxaga et al. (2012)	ASVspoof19-LA	0.46	0.93	0.87	0.14
	ASVspoof19-PA	0.84	0.26	0.88	0.14
RPS Saratxaga et al. (2012)	ASVspoof19-LA	0.52	0.90	0.91	0.10
	ASVspoof19-PA	0.86	0.25	0.89	0.11
SSFC Tapkir et al. (2018)	ASVspoof19-LA	0.50	0.90	0.89	0.12
	ASVspoof19-PA	0.87	0.29	0.90	0.10
GTCC Valero and Alias (2012)	ASVspoof19-LA	0.50	0.90	0.90	0.10
	ASVspoof19-PA	0.84	0.25	0.89	0.12
APGDF Rajan et al. (2013)	ASVspoof19-LA	0.34	0.94	0.82	0.17
	ASVspoof19-PA	0.60	0.17	0.83	0.15

8.2.3 Discussion of the SVM experiments

In these experiments, the CQCC countermeasure performed far better than the other features. In addition, CQCC features continue to work effectively in the presence of microphonic distinctions, where the rest of the spectral features fail. In other results of note, the phase-based APGDF Rajan et al. (2013) features were the second best features compared to the SOTA that achieved the optimal results. Detailed results of the tested SOTA counter measures are presented in Table 8.

Table 9 Experimental performance analysis of the counter measures on ASVspoof2019 and VSDC datasets with a CNN based classifier

Features	Training	Testing	EER	min t-DCF
IMFCC Chakroborty and Saha (2009)	VSDC	VSDC	0.06	0.17
	ASVspoof19	ASVspoof19	0.19	0.50
SSFC Tapkir et al. (2018)	VSDC	VSDC	0.03	0.07
	ASVspoof19	ASVspoof19	0.18	0.46
RFCC Saratxaga et al. (2012)	VSDC	VSDC	0.13	0.25
	ASVspoof19	ASVspoof19	**0.05**	**0.13**
SCFC Kua et al. (2010)	VSDC	VSDC	0.10	0.30
	ASVspoof19	ASVspoof19	0.14	0.36
SCMC Kua et al. (2010)	VSDC	VSDC	0.14	0.37
	ASVspoof19	ASVspoof19	0.44	0.94
PSRCC Tapkir et al. (2018)	VSDC	VSDC	0.08	0.10
	ASVspoof19	ASVspoof19	0.43	0.75
MSRCC Tapkir et al. (2018)	VSDC	VSDC	0.13	0.17
	ASVspoof19	ASVspoof19	0.43	0.75
RPS Saratxaga et al. (2012)	VSDC	VSDC	0.04	0.11
	ASVspoof19	ASVspoof19	0.21	0.65
CQCC Todisco et al. (2017)	VSDC	VSDC	0.04	0.13
	ASVspoof19	ASVspoof19	0.17	0.45
LFCC Zhou et al. (2011)	VSDC	VSDC	0.03	**0.03**
	ASVspoof19	ASVspoof19	0.16	0.43
MFCC Muda et al. (2010)	VSDC	VSDC	0.06	0.16
	ASVspoof19	ASVspoof19	0.26	0.62
LPCC Wong and Sridharan (2001)	VSDC	VSDC	**0.02**	0.05
	ASVspoof19	ASVspoof19	0.19	0.50
GTCC Valero and Alias (2012)	VSDC	VSDC	0.03	0.07
	ASVspoof19	ASVspoof19	0.43	0.75
APGDF Rajan et al. (2013)	VSDC	VSDC	0.01	0.04
	ASVspoof19	ASVspoof19	0.17	0.45

8.3 SOTA features with a CNN classifier

Several deep learning and end-to-end solutions have been published recently to differentiate between spoofed and bonafide speech samples. Therefore, we tested the counter measures against the recent CNN and CNN-GRU based classifiers in order to validate the generalized performance of the highest performing counter measures.

8.3.1 Results on the ASVspoof2019 dataset

In the case of ASVspoof19, the rectangular-based RFCC Saratxaga et al. (2012) features outperform all evaluated counter measures on the ASVspoof19 dataset, with an EER of 0.054% and a min t-DCF of 0.139%. The magnitude-based feature MSRCC was found to be significantly deficient during the ASVspoof19 evaluation, with the highest EER and min

Table 10 Experimental performance analysis of the SOTA counter measures with a CNN-GRU based classifier

Feature	Train dataset	Test dataset	EER	min t-DCF
LFCC Zhou et al. (2011)	VSDC	VSDC	**0.02**	0.05
	ASVspoof19	ASVspoof19	0.15	0.41
SSFC Tapkir et al. (2018)	VSDC	VSDC	0.04	0.07
	ASVspoof19	ASVspoof19	0.17	0.46
MFCC Muda et al. (2010)	VSDC	VSDC	0.03	0.07
	ASVspoof19	ASVspoof19	0.22	0.548
SCFC Kua et al. (2010)	VSDC	VSDC	0.18	0.35
	ASVspoof19	ASVspoof19	0.09	0.25
LPCC Wong and Sridharan (2001)	VSDC	VSDC	0.18	0.03
	ASVspoof19	ASVspoof19	**0.02**	0.47
SCMC Kua et al. (2010)	VSDC	VSDC	0.02	0.05
	ASVspoof19	ASVspoof19	0.17	0.46
MSRCC Tapkir et al. (2018)	VSDC	VSDC	0.21	**0.03**
	ASVspoof19	ASVspoof19	0.22	0.50
PSRCC Tapkir et al. (2018)	VSDC	VSDC	0.23	0.03
	ASVspoof19	ASVspoof19	0.19	0.53
RFCC Saratxaga et al. (2012)	VSDC	VSDC	0.11	0.22
	ASVspoof19	ASVspoof19	0.04	**0.11**
IMFCC Chakroborty and Saha (2009)	VSDC	VSDC	0.03	0.09
	ASVspoof19	ASVspoof19	0.18	0.50
APGDF Rajan et al. (2013)	VSDC	VSDC	0.08	0.26
	ASVspoof19	ASVspoof19	0.14	0.37
CQCC Todisco et al. (2017)	VSDC	VSDC	0.09	0.23
	ASVspoof19	ASVspoof19	0.18	0.49
GTCC Valero and Alias (2012)	VSDC	VSDC	0.04	0.09
	ASVspoof19	ASVspoof19	0.19	0.53
RPS Saratxaga et al. (2012)	VSDC	VSDC	0.04	0.09
	ASVspoof19	ASVspoof19	0.19	0.53

t-DCF values, at 0.443 and 0.947%, respectively. In contrast, the highest EER of 0.21%, and the highest min t-DCF score of 0.43%, were seen in the SCFC Kua et al. (2010) features against VSDC. Detailed results of the SOTA counter measures with the CNN classifier are presented in Table 9.

8.3.2 Results on the VSDC dataset

The results, presented in Table 9 for the VSDC dataset, demonstrate that the APGDF Rajan et al. (2013) front-end features outperform the rest of the studied counter measures, with an EER of 0.01% and a min t-DCF of 0.04%. Following all of the counter measures, For instance, when measured by EER, LPCC Wong and Sridharan (2001) and LFCC Zhou et al. (2011) rank second and third, with scores of 0.026% and 0.032%, respectively.

Table 11 Cross corpus evaluation of the SOTA counter measures with a GMM classifier

Features	Training	Testing	EER	min t-DCF
CQCC Todisco et al. (2017)	VSDC	ASVspoof19	0.40	0.81
	ASVspoof19	VSDC	0.66	0.99
LFCC Zhou et al. (2011)	VSDC	ASV'19 PA	0.60	0.99
	ASV'19 PA	VSDC	0.45	0.99
SSFC Tapkir et al. (2018)	VSDC	ASVspoof19	0.80	0.99
	ASVspoof19	VSDC	0.67	0.99
RPS Saratxaga et al. (2012)	VSDC	ASVspoof19	0.50	0.93
	ASVspoof19	VSDC	0.40	0.94
LPCC Wong and Sridharan (2001)	VSDC	ASVspoof19	0.41	0.89
	ASVspoof19	VSDC	0.62	1.0
PSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	**0.25**	0.39
	ASVspoof19	VSDC	0.67	0.99
MSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	0.29	0.47
	ASVspoof19	VSDC	0.40	0.94
SCFC Kua et al. (2010)	VSDC	ASVspoof19	0.58	0.99
	ASVspoof19	VSDC	0.48	0.99
SCMC Kua et al. (2010)	VSDC	ASVspoof19	0.60	0.97
	ASVspoof19	VSDC	0.46	0.87
MFCC Muda et al. (2010)	VSDC	ASVspoof19	0.42	0.97
	ASVspoof19	VSDC	0.45	0.93
IMFCC Chakroborty and Saha (2009)	VSDC	ASVspoof19	0.45	0.99
	ASVspoof19	VSDC	0.45	0.96
RFCC Saratxaga et al. (2012)	VSDC	ASVspoof19	0.34	0.84
	ASVspoof19	VSDC	**0.39**	**0.27**
GTCC Valero and Alias (2012)	VSDC	ASVspoof19	0.36	**0.38**
	ASVspoof19	VSDC	0.56	1.0
APGDF Rajan et al. (2013)	VSDC	ASVspoof19	0.30	0.59
	ASVspoof19	VSDC	0.61	0.99

8.4 SOTA features with a CNN-GRU classifier

The CNN-GRU classifier is a recent development in the classification of speech samples for voice spoofing detection. This classifier combines the strengths of convolutional neural networks (CNNs) and gated recurrent units (GRUs) to provide a powerful and efficient method for detecting spoofing attacks. The CNN component of the classifier extracts important features from the speech signal, while the GRU component is used to model the temporal dynamics of the speech signal. By combining these two components, the CNN-GRU classifier is able to effectively distinguish between genuine and spoofed speech. Thus, a CNN-GRU Jung et al. (2019) classifier was used to evaluate the SOTA counter measures in this experiment, and the results are presented in Table 10.

Table 12 Cross corpus evaluation of the SOTA counter measures with a CNN based classifier

Feature	Train dataset	Test dataset	EER	min t-DCF
IMFCC Chakroborty and Saha (2009)	VSDC	ASVspoof19	0.43	0.99
	ASVspoof19	VSDC	0.57	1.0
SCMC Kua et al. (2010)	VSDC	ASVspoof19	0.51	0.99
	ASVspoof19	VSDC	**0.35**	0.96
RFCC Saratxaga et al. (2012)	VSDC	ASVspoof19	0.40	0.96
	ASVspoof19	VSDC	0.50	**0.91**
RPS Saratxaga et al. (2012)	VSDC	ASVspoof19	0.68	0.99
	ASVspoof19	VSDC	0.55	0.98
SSFC Tapkir et al. (2018)	VSDC	ASVspoof19	0.72	0.99
	ASVspoof19	VSDC	0.39	**0.91**
PSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	**0.37**	0.94
	ASVspoof19	VSDC	0.51	0.97
MSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	0.46	**0.52**
	ASVspoof19	VSDC	0.47	0.93
CQCC Todisco et al. (2017)	VSDC	ASVspoof19	0.48	0.98
	ASVspoof19	VSDC	0.66	1.0
LFCC Zhou et al. (2011)	VSDC	ASVspoof19	0.51	0.99
	ASVspoof19	VSDC	0.37	0.99
MFCC Muda et al. (2010)	VSDC	ASVspoof19	0.47	0.93
	ASVspoof19	VSDC	0.52	0.94
SCFC Kua et al. (2010)	VSDC	ASVspoof19	0.53	0.99
	ASVspoof19	VSDC	0.54	0.97
LPCC Wong and Sridharan (2001)	VSDC	ASVspoof19	0.46	0.98
	ASVspoof19	VSDC	0.63	1.0
GTCC Valero and Alias (2012)	VSDC	ASVspoof19	0.50	0.98
	ASVspoof19	VSDC	0.41	**0.91**
APGDF Rajan et al. (2013)	VSDC	ASVspoof19	0.44	0.99
	ASVspoof19	VSDC	0.56	1.0

8.4.1 Results on the ASVspoof2019 dataset

When testing of counter measures against the ASVspoof19 dataset, the results demonstrate that LPCC Wong and Sridharan (2001) was the best performing countermeasure, with an EER of 0.02% and a min t-DCF of 0.47%, while the RFCC Saratxaga et al. (2012) and SCFC Kua et al. (2010) counter measures have the lowest EERs at 0.04 and 0.09%, respectively. In contrast, the counter measures with MSRCC and MFCC features obtained the highest EERs of 0.22% respectively. In the case of the min t-DCF, only the RFCC Saratxaga et al. (2012) features performed well, with the lowest value of 0.011%, all other counter measures obtained a min t-DCF score higher than 0.25%, which shows the sensitivity of counter measures in ASV systems.

Table 13 Cross corpus evaluation of the counter measures with a CNN-GRU based classifier

Feature	Train dataset	Test dataset	EER	min t-DCF
LFCC Zhou et al. (2011)	VSDC	ASVspoof19	0.47	0.99
	ASVspoof19	VSDC	0.53	0.98
MFCC Muda et al. (2010)	VSDC	ASVspoof19	0.45	0.94
	ASVspoof19	VSDC	**0.34**	0.97
SCFC Kua et al. (2010)	VSDC	ASVspoof19	0.49	0.99
	ASVspoof19	VSDC	0.59	0.99
LPCC Wong and Sridharan (2001)	VSDC	ASVspoof19	0.65	0.99
	ASVspoof19	VSDC	0.45	1.0
SCMC Kua et al. (2010)	VSDC	ASVspoof19	0.45	0.99
	ASVspoof19	VSDC	0.65	1.0
PSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	0.52	0.98
	ASVspoof19	VSDC	0.57	0.99
MSRCC Tapkir et al. (2018)	VSDC	ASVspoof19	0.49	0.99
	ASVspoof19	VSDC	0.54	1.0
RFCC Saratxaga et al. (2012)	VSDC	ASVspoof19	0.44	0.99
	ASVspoof19	VSDC	0.67	1.0
IMFCC Chakroborty and Saha (2009)	VSDC	ASVspoof19	0.44	0.99
	ASVspoof19	VSDC	0.61	1.0
APGDF Rajan et al. (2013)	VSDC	ASVspoof19	**0.34**	**0.82**
	ASVspoof19	VSDC	0.51	**0.96**
CQCC Todisco et al. (2017)	VSDC	ASVspoof19	0.50	0.99
	ASVspoof19	VSDC	0.43	1.0
GTCC Valero and Alias (2012)	VSDC	ASVspoof19	0.49	0.99
	ASVspoof19	VSDC	0.54	1.0
SSFC Tapkir et al. (2018)	VSDC	ASVspoof19	0.51	0.99
	ASVspoof19	VSDC	0.56	1.0
RPS Saratxaga et al. (2012)	VSDC	ASVspoof19	0.50	0.96
	ASVspoof19	VSDC	0.55	0.98

Table 14 Computational details of the SOTA counter measures

Model	ASVspoof2019		ASVspoof2021		VSDC
	EER	t-DCF	EER	t-DCF	EER
ASSIST Jung et al. (2022)	**0.83**	**0.027**	15.84	0.537	51.21
ASSIST-L Jung et al. (2022)	0.99	0.030	**15.69**	0.520	59.32
RawBoost Tak et al. (2021)	2.17	0.055	23.1	0.759	64.90
One class Zhang et al. (2021)	5.38	0.309	48.7	0.901	37.95
Res2Net34 Li et al. (2021)	1.67	0.058	48.00	0.994	29.00
Res2Net50 Li et al. (2021)	1.63	0.049	39.00	0.999	**24.00**
ResNet34-LA Aravind et al. (2020)	10.53	0.19	26.55	**0.34**	40.23
ResNet34-PA Aravind et al. (2020)	6.61	0.15	49.51	0.41	34.12

8.4.2 Results on the VSDC dataset

The results from Table 10 demonstrate that in the case of the VSDC dataset, the linear and spectral magnitude-based LFCC and SCMC features achieved the same performance, with an EER of 0.02% and a min t-DCF of 0.05%. Due to the versatility of the VSDC dataset, the SCFC features scored an EER of 0.18% and a min t-DCF of 0.35%. In particular, the results from this experiment proved to be optimal for the VSDC dataset, in comparison with the GMM, SVM, and CNN-GRU classifiers.

8.4.3 Discussion of the CNN/CNN-GRU results

These experiments clearly show the significance of the choice of back-end classifiers in speech spoofing detection. For instance, the worst EER recorded with the GMM classifier was 0.80% on the LFCC feature, while the comparable worst EER achieved with the CNN classifier was just 0.265% with MFCC Muda et al. (2010), i.e., when using the same feature set but replacing the classifier, the worst case EER was reduced by nearly 65%. Linear-based features perform well with the CNN-based classifier to detect voice spoofing attacks, as seen in LFCC Zhou et al. (2011) and LPCC Wong and Sridharan (2001) in these experiments.

8.5 Discussion of the overall performance of counter measures against all four classifiers

Based on the experiments involving the SOTA counter measures with machine learning-based GMM and SVM-based classifiers, CQCC Todisco et al. (2017) is the best general countermeasure against voice spoofing attacks. More specifically, in the testing of replay attacks, where the performance of the SOTA classifiers drops significantly, the CQCC features scored the best EER, 0.27, 0.43, and 0.30% in GMM and SVM testing, respectively. The results of both ML-based classifiers show that the phase-based APGDF Rajan et al. (2013), linear LFCC Zhou et al. (2011), and inverted Mel-based IMFCC Chakroborty and Saha (2009) counter measures performed better than the comparative methods in all experiments and received the optimal scores in all performance measurements. APGDF also had the lowest EER and min t-DCF in the ASVspoof2019-PA dataset evaluation. Similar results were obtained with VSDC on the SVM classifier. In comparison to the comparative approaches, the SCFC Kua et al. (2010) features failed to perform well in any experiment, and had the highest EER and min t-DCF, in fact, the performance of the majority of the spectral and cepstral counter measures declined when tested against other classifiers in different configurations.

When compared to the GMM classifier, the CNN classifier performed significantly better across the board. In addition, the CNN-GRU classifier proved to be the optimal classifier for both datasets. Therefore, it may be concluded from these experiments that the performance of the existing SOTA countermeasure differs in the presence of distinct classifiers and may not function adequately in the presence of advanced spoofing attempts. However, the effectiveness of the SOTA countermeasure across corpora must still be demonstrated. To this end, in the next section we provide a cross-corpus evaluation of the SOTA counter measures.

8.6 Experimental analysis of handcrafted counter measures across corpora

One of the shortfalls in the current research into audio spoof detection is that cross-corpus evaluations are rarely done, limiting knowledge of the generalizability of both features and full countermeasure frameworks. To overcome this knowledge gap, this comparative analysis performs a cross-corpus evaluation, where one dataset is used for training and another for evaluation, to show the effectiveness of these features at generalized spoof detection across different conditions present in the respective datasets. The results of the cross corpus evaluations against GMM, CNN and CNN-GRU are presented in Tables 11, 12 and 13.

In the instance of VSDC training and ASVspoof19 testing, the APGDF-based countermeasure Rajan et al. (2013) obtained the best overall EER of these experiments, 0.34%, and a min t-DCF score of 0.82%. With an EER of 0.44%, the RFCC Saratxaga et al. (2012) and IMFCC Chakroborty and Saha (2009) were ranked second and third, respectively. In the case of ASVspoof19 training and VSDC testing, the MFCC-based countermeasure Muda et al. (2010) outperformed the SOTA features, with the lowest EER of 0.34% and the lowest min t-DCF score of 0.97%. The outcomes of this experiment were similar to the results of the CNN-based cross-corpus evaluations. The min t-DCF of the SOTA counter measures was significantly greater than in standalone testing.

In the instance of ASVspoof19 dataset training and VSDC testing, the results demonstrate that in the majority of the counter measures, the min t-DCF was at or around 1%. These findings demonstrate the ineffectiveness of existing counter measures and classifiers against unknown spoofing attacks. Furthermore, these findings have raised concerns regarding the effectiveness of the existing protections against complicated contemporary spoofing attacks.

8.7 Experimental analysis of recent end-to-end counter measures across corpora

Following the observation of a performance drop in SOTA counter measures we conducted a cross-corpus sub-experiment utilizing contemporary deep learning and end-to-end solutions, and the performance of recent deep learning and end-to-end solutions is shown in Table 14.

Several deep learning and end-to-end solutions have demonstrated outstanding performance in the recent ASV challenges Jung et al. (2022); Tak et al. (2021); Zhang et al. (2021); Aravind et al. (2020); Li et al. (2021). In the latest competition, ASVspoof2021, the end-to-end solution ASSIST Jung et al. (2022) achieved the best EER and min t-DCF on the ASVspoof19-LA data subset. Similarly, Zhang et al. (2021) and Tak et al. (2021) also performed well in this task. Many participants, however, presented deep learning and end-to-end systems that had been trained and tested to mitigate only a single attack. To demonstrate the general effectiveness of these counter measures we performed an experiment in which we evaluated the higher performing solutions to the ASV challenges across corpora. Each solution was trained and tested on the ASVspoof2019-LA subset and evaluated against the LA speech samples of the ASVspoof2021 corpus. Similarly, if the model's stated results were based on ASVspoof2021, we examined that model's performance against ASVspoof2019. The key reasoning behind this experiment was to evaluate and report the generalizability of these counter measures.

To measure generalizability, we selected the top-performing models with publicly available code. It was observed from the results that the top performer, ASSIST Jung et al.

(2022), and its companion model ASSIST-L, achieved EER values of 0.83 and 0.99%, respectively, against the ASVspoof2019 dataset. However, the performance of these models deteriorated when tested against speech samples from the ASVspoof2021 dataset. Specifically, the EER scores increased to 15.84 and 15.69%, and the min t-DCFs rose from 0.027 to 0.537%, and from 0.030 to 0.520%, respectively. Furthermore, when tested against ASV spoof 2021, the EER for the solution proposed in Zhang et al. (2021) increased from 2.17 to 23.1%, and the min t-DCF increases from 0.055 to 0.759%, however the worst performance was observed in the case of the RawBoost solution proposed in Tak et al. (2021). Although RawBoost performed optimally using augmented data, with an EER score of 2.17%, its EER and min t-DCF increased dramatically when tested on other datasets and without augmentation. In the case of ASVspoof2021 and without data augmentation, the EER of Tak et al. (2021) increased by 43% and min t-DCF increased from 0.301 to 0.903%.

Although the adaptation of the Resnet-34 architecture published in Aravind et al. (2020) reported an EER of 5.32% against the ASVSpoof2019-LA dataset, and 5.74% against the ASVSpoof2019-PA, these results could not be replicated. We observed EERs of 10.53 and 6.61%, respectively, in our configured environment. When tested on ASVSpoof2021, performance dropped even further, with EER results of 26.55% on LA and 49.51% on PA.

To further test the extensibility of this and other methods, we performed cross-attack, cross-corpus testing, with training on ASVspoof2019-PA and testing against ASVSpoof2021-deepfake (DF) dataset, which achieved an EER 42.21%. Similar performance deterioration was observed for Aravind et al. (2020) on the distinct dataset, where EER rose to 48.00% and 39.00% for res2net34 and res2net50, respectively.

8.8 Overall performance of the SOTA voice spoofing solutions, and a way forward

Based on the results of this study's experiments, we can infer that the none of the selected counter measures was consistent over a wide range of datasets and classifiers. As shown in the preceding subsections, several counter measures, such as CQCC Todisco et al. (2017) and RFCC Saratxaga et al. (2012), performed better overall in standalone testing, however, when the dataset and classifier were changed, these counter measures failed to perform well. While PSRCC Tapkir et al. (2018) and MSRCC Tapkir et al. (2018) did not perform much better in standalone testing, they did significantly better in cross-corpus examination. After extensive testing, it is extremely difficult to select the best countermeasure from among those tested because some performed well with one data set but failed at another. We may infer from this that classifiers and datasets have a significant influence on the present counter measures. In addition, in standalone corpus testing with four alternative classifiers, CQCC Todisco et al. (2017) performed consistently with only a minor decline in performance across classifiers. PSRCC Tapkir et al. (2018) and APGDF Rajan et al. (2013) scored marginally better when the classifier varied and in cross-corpus evaluation. As a consequence of the extensive testing with various datasets and classifiers, the effectiveness of the existing SOTA counter measures was called into question. These tests also demonstrated the significance of cross-corpus and variable classifier evaluation of proposed solutions.

The scores achieved for the end-to-end solutions imply that while these methods perform well on a specific dataset the results do not transfer well. These tests demonstrate the impact of formulating a solution based on a particular corpus, as the performance of deep learning and end-to-end solutions degrades significantly when presented with unknowns. Each static dataset differs and has distinct discrepancies due to varied factors, such as the

ambient setting, unknown equipment, the influence of microphone type, and the spoofing technique used to produce spoof samples. Therefore, developing a spoofing countermeasure without taking into consideration the aforementioned factors may lead to performance degradation when tested across corpora.

Thus, based on these results, we conclude that there is a dire need to concentrate on a broad approach that mitigates the aforementioned variances while developing robust voice spoofing defenses, with an accompanying shift in the research community's focus to environmental or scenario-based spoofing counter measures. The environment is an essential part of spoofing detection counter measures; for instance, several counter measures may fail to provide adequate performance when surrounded by a loud environment or background noise. As a result, a thorough comparison study is required to find the best countermeasure for specific scenarios.

In addition, the need for cross dataset evaluation is important to consider when evaluating the performance of any aspect of an audio spoof detection system, as shown in this paper. Only when counter measures are evaluated across corpora will important insights into the robustness of the tested features be revealed. These may provide additional information and direction toward a more general and complete ASV system.

8.9 Limitations and future work

Aside from the effectiveness of the presented work, there are certain limitations to this study. Due to contradictory technical issues and lack of code availability, only fourteen counter measures were chosen to do the cross-corpus and multi-classifier evaluation. Moreover, in accordance with the advancement of the voice spoofing domain and deep learning methodologies such as CNNs, transformers, and so on, we compared four recent top-rated (by EER and min t-DCF scores on ASVspoof datasets) deep learning and end-to-end solutions with cross corpus evaluation. Despite the fact that just eight models were examined, the findings motivate us to continue working in the deep learning and end-to-end arenas. Thus, we want to expand this work in the future to include the most recent deep learning/end-to-end solutions in order to provide an apples-to-apples comparison. Lastly, we intend to create a benchmark for comparing CNN and DNN-based architectures for the automatic speaker verification system.

9 Existing issues and future needs

This section describes the challenges and issues observed during this review as well as the future requirements for developing dependable and secure anti-spoofing solutions.

9.1 Need for explainable AI in audio forensics

Explainable AI (XAI) is an emerging field which emphasises the need for human understanding through what has traditionally been a black box process. An important gain over the interaction between humans and XAI is a level of trust in the process, which studies have shown does not currently exist Gerlings et al. (2020). Existing deepfake detection approaches are typically designed to perform batch analysis over a large dataset, however, when these techniques are employed in the field by journalists or law enforcement, there may only be a small set of videos available for analysis. A numerical score parallel to the

probability of an audio or video being real or fake is not valuable to these users if it cannot be confirmed with an appropriate proof of the score. If results for single interactions between humans and computers are to be accepted it is vitally important to involve XAI early in the process.

9.2 Fairness-enabled voice spoofing counter measures and ASV

Fairness is defined as the lack of bias or preference toward a person or a group based on intrinsic or acquired attributes. Prior to their widespread commercialization, it is necessary to study the bias and impartiality of spoofing detection and speaker verification systems across populations. A quantitative evaluation has been done to study fairness with respect to gender, ethnicity, or other factors, in the ASV domain. To avoid the real-world consequences of a biased and flawed system that favors a single sub-group or class, it is vital to ensure the impartiality of data sets used for training and testing spoofing systems.

9.3 The role of adversarial machine learning on existing audio spoofing counter measures

High-performance spoofing countermeasure systems for automated speaker verification (ASV) have been presented during the ASVspoof sessions. Adversarial attacks pose a significant threat to ASV systems and their counter measures, however, the robustness of these systems in the face of adversarial attacks has yet to be thoroughly examined. Currently, there are only a limited number of counter measures that address this domain. It is therefore necessary to develop a more adversarial-aware ASV system.

9.4 Need For liveliness detection in voice spoofing detection

Liveliness detection is a crucial component in voice spoofing detection systems as it verifies the authenticity of the speech being generated. As technology advances, spoofing attacks using recorded or synthetic speech are becoming increasingly prevalent. Thus, it is important to ensure that the speech being analyzed is not just a recording but is generated by a live person in real-time. This can be achieved through the verification of certain physiological and behavioral characteristics that are unique to live speech, such as the presence of breath, heartbeats, or other subtle movements. By integrating liveliness detection into voice spoofing detection systems, a more comprehensive approach to securing voice-based systems may be achieved, as it provides an additional layer of security. Furthermore, liveliness detection can also help reduce false rejections in voice spoofing detection systems, thereby improving the overall user experience in high-security situations such as financial transactions and access control.

9.5 Robust PAD for ASV in smart homes

Because the Smart Home concept encourages hands-free automation, bio-metric technologies, and particularly automatic speech recognition, are the ideal way to regulate

and personalize access. However, while contemporary spoofing systems are extremely descriptive and specific when detecting presentation attacks, current speaker identification systems may be vulnerable. Therefore, most of these PAD systems are ineffective in the current environment where the types of presentation attacks are unknown. When deployed to smart homes, the risk of attack increases significantly. Hence, there is an immediate need to build a strong PAD system to protect ASV systems in order to accelerate the safe use of ASVs in Smart Home applications.

9.6 Need for new audio forensics datasets

Voice-controlled devices (VCDs), aided by the Internet of Things (IoT), have enabled the development, personalization, and automation, of smart homes through voice activation. Unfortunately, in many cases these VCDs are largely or completely unprotected and can be exploited via spoofing attacks. Existing datasets, such as the Voice spoofing detection corpus (VSDC), ASVspoof 2017, ASVspoof 2019, and ReMASC, contain large-scale replay (i.e., single and multi-order replays) and basic cloning audio files, but newer, deep learning-based attacks easily defeat current solutions, introducing a new level of complexity for anti-spoofing algorithms. Moreover, large-scale datasets like ASVspoof 2017 and ASVspoof 2019 lack cloning design parameters, which are an essential component of current VCDs. Therefore, an enhanced cloning dataset is needed in order to stay ahead of the curve of advanced voice cloning spoof attacks. For creation of such dataset, new generative algorithms such as diffusion models could be employed.

9.7 Evaluation of ASV performance on non-English datasets

Although English has become the *de facto* standard for the evaluation of voice spoofing techniques and counter measures, care must be taken to avoid grouping all languages together for this purpose. Each discrete language comes with a unique set of challenges. For instance, Almutairi and Elgibreen (2022) points out that mispronunciation of Arabic's vowel sounds can completely change the meaning of an entire sentence, something not very prevalent in English. Existing methods cannot be trusted when presented with foreign languages. When state of the art methods, trained on English datasets, were evaluated on non-English data, Ba et al. (2023) found models degraded as much as 300%. There is therefore a need for extensive study in this area.

9.8 Need for continued cross-dataset evaluation

The need for cross dataset evaluation is important to consider when evaluating the performance of any aspect of an audio spoof detection system as shown in this paper. Cross dataset evaluation gives important insight into the robustness of the tested features and can provide additional information and direction toward a more general and complete ASV system. It is therefore vitally important that evaluation across corpora be a focus in continuing research in this domain.

9.9 Attack-aware audio forensics counter measures

ASV systems should always discard both non-target, e.g., speakers not authorized on the system, and spoofed, i.e., reprocessed or transformed, signals. However, little consideration has been given to how ASV systems should be changed when they encounter spoofing attacks, or even when they pair up with spoofing counter measures, much less to how both systems might be jointly optimized.

9.10 Cross-lingual and cross-cultural

Cross-lingual and cross-cultural voice spoofing detection has become a crucial need in today's world, as the use of voice recognition technology is rapidly increasing. The ability of a voice spoofing system to detect fake voice inputs in different languages, and with respect to cultural differences, is essential to ensure the reliability and accuracy of the system. To take it a step further, the globalization of technology has made it possible for individuals with malicious intent to attempt to bypass security systems by using different languages and accents. Therefore, the development of cross-lingual and cross-cultural voice spoofing detection systems is necessary to provide a robust security layer and prevent potential threats.

9.11 Need for novel features and counter measures for addressing compression and encoding-based artifacts in ASV systems

The focus of ASVspoof2021, the most recent ASV challenge, emphasizes the importance of voice spoofing counter measures, especially in terms of compression and encoding-based artifacts, which can have a major influence on an ASV system's performance and limit its capacity to accurately identify spoofed speech. This highlights the importance of ongoing research and the need for improvement in the field of speech spoofing detection. As attacking tactics for exploiting ASV systems get more complex and creative, researchers must rise to meet the challenge. To reach this goal and increase the resilience of ASV systems against spoofing attacks, novel features and counter measures specially tailored to address the issues introduced by compression and encoding-based artifacts are required.

9.12 Physics, domain knowledge, and AI

Physics-informed neural networks (PINNs) are becoming increasingly popular due to their ability to incorporate existing knowledge of physics into deep learning models, making them ideal for applications where the underlying physics is known Raissi et al. (2019). This is especially relevant to voice spoofing detection, where signal processing properties of the human voice must be fully understood and modeled in order to discriminate between real and fake speech. PINNs can also adapt to uncertainty and variability in the data, which is a critical aspect in voice spoofing detection, as real-world speech signals are often noisy and dynamic. Because PINNs can include both physical and statistical models in a single framework, they may well be an effective tool for the detection of spoofed speech and are projected to generate more accurate and robust results than traditional machine learning.

9.13 Challenges, issues, and research needs for integrated ASV systems

Integrated solutions that combine speech spoofing detection and automated speaker verification have a number of associated impediments and pitfalls. The trade-off between false acceptance and false rejection rates is a key challenge, as the system must balance the need to correctly validate real speakers with the necessity to recognize and reject spoofing attempts. Also, such systems require computationally complex operations and may not be suitable to application on resource-poor systems. Another concern is the unpredictability of speech patterns across different environments, dialects, and recording conditions, which can have a detrimental influence on system performance. Furthermore, the rapid development of improved speech spoofing techniques has made it more difficult to identify spoofing attempts, emphasizing the significance of ongoing research and development in this field. These limits and constraints emphasize the importance of continued research to improve the resilience and dependability of integrated ASV systems and make them more resilient.

10 Conclusion

This research provides a comprehensive review of voice spoofing attacks, counter measures proposed to counter certain attacks, publicly accessible datasets, and the criteria used to evaluate countermeasure performance in voice spoofing detection. It also includes an experimental comparison of the SOTA (SOTA) voice spoofing counter measures against multi-and cross-datasets and classifiers. Based on the results of the experiments, this study concludes that the characteristics that better capture microphone signatures and harmonic distortions give improved detection performance for the identification of PA attacks. Furthermore, we show that the efficacy of SOTA countermeasure varies significantly with changes to the dataset and classifiers. Extensive testing reveals that although these counter measures perform well against known attacks on a specific dataset there is a dire need for a robust and generalize CM. Detailed experiments also demonstrate the significance of cross-corpus evaluation for future voice spoofing technologies.

11 Reproducibility

We have provided a GitHub repository[2] that contains all the code and explanation needed to reproduce results of models evaluated to test their generalizability. The README.pdf file contains explanation of how to run the experiments.

Acknowledgements This material is based upon work supported by the National Science Foundation (NSF) under award number 1815724 and Michigan Transnationals Research and Commercialization (MTRAC), Advanced Computing Technologies (ACT) award number 292883. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and MTRAC ACT.

[2] https://github.com/smileslab/Comparative-Analysis-Voice-Spoofing.

References

Alegre F, Amehraye A, Evans N (2013) Spoofing counter measures to protect automatic speaker verification from voice conversion. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp. 3068–3072. IEEE

Aljasem M, Irtaza A, Malik H, Saba N, Javed A, Malik KM, Meharmohammadi M (2021) Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging. IEEE Trans Inf Forensics Secur 16:3524–3537. https://doi.org/10.1109/TIFS.2021.3082303

Almutairi Z, Elgibreen H (2022) A review of modern audio deepfake detection methods: challenges and future directions. Algorithms 15(5):155

Aravind P, Nechiyil U, Paramparambath N, et al (2020) Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:2008.03464

Arif T, Javed A, Alhameed M, Jeribi F, Tahir A (2021) Voice spoofing countermeasure for logical access attacks detection. IEEE Access 9:162857–162868. https://doi.org/10.1109/ACCESS.2021.3133134

Ba Z, Wen Q, Cheng P, Wang Y, Lin F, Lu L, Liu Z (2023) Transferring audio deepfake detection capability across languages. In: Proceedings of the ACM web conference 2023, pp. 2033–2044

Balamurali B, Lin KE, Lui S, Chen J-M, Herremans D (2019) Toward robust audio spoofing detection: a detailed comparison of traditional and learned features. IEEE Access 7:84229–84241

Balamurali BT, Lin KE, Lui S, Chen J-M, Herremans D (2019) Toward robust audio spoofing detection: a detailed comparison of traditional and learned features. IEEE Access 7:84229–84241. https://doi.org/10.1109/ACCESS.2019.2923806

Baumann R, Malik KM, Javed A, Ball A, Kujawa B, Malik H (2021) Voice spoofing detection corpus for single and multi-order audio replays. Comput Speech Lang 65:101132

Cai W, Cai D, Liu W, Li G, Li M (2017) counter measures for automatic speaker verification replay spoofing attack: on data augmentation, feature representation, classification and fusion. In INTERSPEECH, pp. 17–21

Cai W, Doshi A, Valle R (2018) Attacking speaker recognition with deep generative models. CoRR **abs/1801.02384**arXiv:1801.02384

Chakroborty S, Saha G (2009) Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter. Int J Signal Process 5(1):11–19

Chen N, Qian Y, Dinkel H, Chen B, Yu K (2015) Robust deep feature for spoofing detection–the sjtu system for asvspoof 2015 challenge. In: Sixteenth annual conference of the international speech communication association

Chen Z, Xie Z, Zhang W, Xu X (2017) ResNet and model fusion for automatic spoofing detection. Interspeech 2017: 102–106. https://doi.org/10.21437/Interspeech.2017-1085

Chen X, Zhang Y, Zhu G, Duan Z (2021) UR channel-robust synthetic speech detection system for ASVspoof 2021. In: Proc. 2021 edition of the automatic speaker verification and spoofing counter measures challenge, pp. 75–82. https://doi.org/10.21437/ASVSPOOF.2021-12

Chen F, Deng S, Zheng T, He Y, Han J (2023) Graph-based spectro-temporal dependency modeling for anti-spoofing. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096741

Chettri B, Kinnunen T, Benetos E (2020) Deep generative variational autoencoding for replay spoof detection in automatic speaker verification

Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622

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

Delgado H, Evans N, Kinnunen T, Lee KA, Liu X, Nautsch A, Patino J, Sahidullah M, Todisco M, Wang X, et al (2021) Asvspoof 2021: automatic speaker verification spoofing and counter measures challenge evaluation plan. arXiv preprint arXiv:2109.00535

Ding S, Zhang Y, Duan Z (2023) Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094704

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), pp. 4860–4864

Ergünay SK, Khoury E, Lazaridis A, Marcel S (2015) On the vulnerability of speaker verification to realistic voice spoofing. In: 2015 IEEE 7th international conference on biometrics theory, applications and systems (BTAS), pp. 1–6. IEEE

Font R, Espín JM, Cano MJ (2017) Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge. In: Interspeech, pp. 7–11

Gerlings J, Shollo A, Constantiou I (2020) Reviewing the need for explainable artificial iintelligence (xAI). arXiv. https://doi.org/10.48550/ARXIV.2012.01007. arXiv:2012.01007

grandviewresearch: voice biometrics market forecast. https://www.grandviewresearch.com/industry-analysis/us-voice-recognition-market. Accessed: May 25 2023

Gunendradasan T, Wickramasinghe B, Le P, Ambikairajah E, Epps J (2018) Detection of replay-spoofing attacks using frequency modulation features. In INTERSPEECH, pp. 636–640. https://doi.org/10.21437/Interspeech.2018-1473

Gunendradasan T, Irtza S, Ambikairajah E, Epps J (2019) Transmission line cochlear model based am-fm features for replay attack detection. In: ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6136–6140. https://doi.org/10.1109/ICASSP.2019.8682771

Hassan F, Javed A (2021) Voice spoofing countermeasure for synthetic speech detection. In: 2021 International conference on artificial intelligence (ICAI), pp. 209–212. https://doi.org/10.1109/ICAI52203.2021.9445238

Huang L, Pun C-M (2020) Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced densenet-bilstm network. IEEE/ACM Trans Audio Speech Lang Process 28:1813–1825

Jati A, Hsu C-C, Pal M, Peri R, AbdAlmageed W, Narayanan S (2021) Adversarial attack and defense strategies for deep speaker recognition systems. Comput Speech Lang 68:101199. https://doi.org/10.1016/j.csl.2021.101199

Javed A, Malik KM, Irtaza A, Malik H (2021) Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. Appl Acoust 183:108283

Javed A, Malik KM, Malik H, Irtaza A (2022) Voice spoofing detector: a unified anti-spoofing framework. Expert Syst Appl 198:116770

Ji Z, Li Z-Y, Li P, An M, Gao S, Wu D, Zhao F (2017) Ensemble learning for countermeasure of audio replay spoofing attack in asvspoof2017. In: Interspeech, pp. 87–91

Jose A, Joseph J, Devadhas G, Shinu MM (2018) Influence of filter bank structure on the statistical significance of coefficients in cepstral analysis for acoustic signals. In: Thampi, S.M., Krishnan, S.r., Corchado Rodriguez, J.M., Das, S., Wozniak, M., Al-Jumeily, D. (eds.) Advances in signal processing and intelligent recognition systems, Springer, Cham, pp. 91–104

Jung J-w, Shim H-j, Heo H-S, Yu H-J (2019) Replay attack detection with complementary high-resolution information using end-to-end dnn for the asvspoof 2019 challenge. arXiv preprint arXiv:1904.10134

Jung J-w, Heo H-S, Tak H, Shim H-j, Chung JS, Lee B-J, Yu H-J, Evans N (2021) AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. arXiv . https://doi.org/10.48550/ARXIV.2110.01200. arXiv:2110.01200

Jung J-w, Heo H-S, Tak H, Shim H-j, Chung JS, Lee B-J, Yu H-J, Evans N (2022) Aasist: audio anti-spoofing using integrated spectro-temporal graph attention networks. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6367–6371. IEEE

Jung J-w, Tak H, Shim H-j, Heo H-S, Lee B-J, Chung S-W, Yu H-J, Evans N, Kinnunen T (2022) Sasv 2022: The first spoofing-aware speaker verification challenge. arXiv preprint arXiv:2203.14732

Kamble MR, Sailor HB, Patil HA, Li H (2020) Advances in anti-spoofing: from the perspective of asvspoof challenges. APSIPA Trans Signal and Inf Process 9:e2

Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee KA (2017a)The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: INTERSPEECH. https://doi.org/10.21437/Interspeech.2017-1111

Kinnunen TH, Sahidullah M, Falcone M, Costantini L, Hautamäki RG, Thomsen DAL, Sarkar AK, Tan Z, Delgado H, Todisco M, Evans NWD, Hautamäki V, Lee K-A (2017b) Reddots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research. 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 5395–5399

Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee KA (2017c) The asvspoof 2017 challenge: assessing the limits of replay spoofing attack detection

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 counter measures and automatic speaker verification. arXiv preprint arXiv:1804.09618

Korshunov P, Marcel S, Muckenhirn H, Gonçalves AR, Mello AS, Violato RV, Simoes FO, Neto MU, de Assis Angeloni M, Stuchi JA, et al (2016) Overview of btas 2016 speaker anti-spoofing competition. In: 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS), pp. 1–6. IEEE

Korshunov P, Gonçalves AR, Violato RP, Simões FO, Marcel S (2018) On the use of convolutional neural networks for speech presentation attack detection. In: 2018 IEEE 4th international conference on identity, security, and behavior analysis (ISBA), pp. 1–8. IEEE

Kreuk F, Adi Y, Cisse M, Keshet J (2018) Fooling end-to-end speaker verification with adversarial examples. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1962–1966. IEEE

Kua JMK, Thiruvaran T, Nosratighods M, Ambikairajah E, Epps J (2010) Investigation of spectral centroid magnitude and frequency for speaker recognition. In: Odyssey, p. 7

Kwak I-Y, Kwag S, Lee J, Jeon Y, Hwang J, Choi H-J, Yang J-H, Han S-Y, Huh JH, Lee C-H, Yoon JW (2023) Voice spoofing detection through residual network, max feature map, and depthwise separable convolution. IEEE Access 11:49140–49152. https://doi.org/10.1109/ACCESS.2023.3275790

Kwon H, Lee K, Ryu J, Lee J (2023) Audio adversarial example detection using the audio style transfer learning method. IEEE Access

Lai C-I, Abad A, Richmond K, Yamagishi J, Dehak N, King S (2019) Attentive filtering networks for audio replay attack detection. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6316–6320. IEEE

Lai C-I, Chen N, Villalba J, Dehak N (2019) Assert: Anti-spoofing with squeeze-excitation and residual networks. arXiv preprint arXiv:1904.01120

Lee K-A, Larcher A, Wang G, Kenny P, Brümmer N, van Leeuwen DA, Aronowitz H, Kockmann M, Vaquero C, Ma B, Li H, Stafylakis T, Alam MJ, Swart A, Pérez J (2015) The reddots data collection for speaker recognition. In INTERSPEECH

Li Z, Shi C, Xie Y, Liu J, Yuan B, Chen Y (2020) Practical adversarial attacks against speaker recognition systems. In: Proceedings of the 21st international workshop on mobile computing systems and applications, pp. 9–14

Li X, Li N, Weng C, Liu X, Su D, Yu D, Meng H (2021) Replay and synthetic speech detection with res2net architecture. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6354–6358. IEEE

Liu L, Yang J (2020) Study on feature complementarity of statistics, energy, and principal information for spoofing detection. IEEE Access 8:141170–141181. https://doi.org/10.1109/ACCESS.2020.3013066

Liu X, Sahidullah M, Kinnunen T (2022) Spoofing-aware speaker verification with unsupervised domain adaptation. arXiv preprint arXiv:2203.10992

Ma Y, Ren Z, Xu S (2021) RW-Resnet: a novel speech anti-spoofing model using raw waveform. arXiv. https://doi.org/10.48550/ARXIV.2108.05684. arXiv:2108.05684

Malik KM, Javed A, Malik H, Irtaza A (2020) A light-weight replay detection framework for voice controlled IoT devices. IEEE J Sel Top Signal Processing 14(5):982–996

Malik KM, Javed A, Malik H, Irtaza A (2020) A light-weight replay detection framework for voice controlled IoT devices. IEEE J Sel Top Signal Process 14(5):982–996. https://doi.org/10.1109/JSTSP.2020.2999828

Mittal A, Dua M (2021) Automatic speaker verification systems and spoof detection techniques: review and analysis. Int J Speech Technol 25:1–30

Muda L, Begam M, Elamvazuthi I (2010) Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. arXiv preprint arXiv:1003.4083

Nagarsheth P, el Khoury E, Patil K, Garland M (2017) Replay attack detection using dnn for channel discrimination. In INTERSPEECH

Nagrani A, Chung JS, Zisserman A (2017) Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612

Naika R (2018) An overview of automatic speaker verification system. Intelligent computing and information and communication. Springer, Cham, pp 603–610

Novoselov S, Kozlov A, Lavrentyeva G, Simonchik K, Shchemelinin V (2016) Stc anti-spoofing systems for the asvspoof 2015 challenge. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5475–5479. IEEE

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), pp. 1047–1053. IEEE

Paul D, Pal M, Saha G (2015) Novel speech features for improved detection of spoofing attacks. In: 2015 annual IEEE India conference (INDICON), pp. 1–6. IEEE

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Raissi M, Perdikaris P, Karniadakis GE (2019) Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput phys 378:686–707

Rajan P, Kinnunen T, Hanilci C, Pohjalainen J, Alku P (2013) Using group delay functions from all-pole models for speaker recognition. In INTERSPEECH, pp. 2489–2493. Citeseer

Research EM (2023) Voice biometrics market forecast. https://www.expertmarketresearch.com/reports/voice-assistant-application-market. Accessed: May 25 2023

Rostami AM, Homayounpour MM, Nickabadi A (2021) Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. arXiv preprint arXiv:2109.02051

Sahidullah M, Delgado H, Todisco M, Yu H, Kinnunen T, Evans N, Tan Z-H (2016) Integrated spoofing counter measures and automatic speaker verification: an evaluation on asvspoof 2015

Sahidullah M, Delgado H, Todisco M, Kinnunen T, Evans N, Yamagishi J, Lee K-A (2019) Introduction to voice presentation attack detection and recent advances. Handbook of biometric anti-spoofing. Springer, New York, pp 321–361

Saito Y, Takamichi S, Saruwatari H (2018) Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans Audio Speech Lang Process 26(1):84–96. https://doi.org/10.1109/TASLP.2017.2761547

Saranya MS, Padmanabhan R, Murthy HA (2018) Replay attack detection in speaker verification using non-voiced segments and decision level feature switching. In: 2018 International conference on signal processing and communications (SPCOM), pp. 332–336. https://doi.org/10.1109/SPCOM.2018.8724469

Saratxaga I, Hernáez I, Pucher M, Sainz I (2012) Perceptual importance of the phase related information in speech. In: INTERSPEECH, vol. 2. https://doi.org/10.21437/Interspeech.2012-411

Suthokumar G, Sriskandaraja K, Sethu V, Wijenayake C, Ambikairajah E (2019) Phoneme specific modelling and scoring techniques for anti spoofing system. In: ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6106–6110. https://doi.org/10.1109/ICASSP.2019.8682411

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks

Tak H, Kamble M, Patino J, Todisco M, Evans N (2021) RawBoost: a raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. arXiv. https://doi.org/10.48550/ARXIV.2111.04433. arXiv:2111.04433

Tak H, Jung J-w, Patino J, Kamble M, Todisco M, Evans N (2021) End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. arXiv. https://doi.org/10.48550/ARXIV.2107.12710. arXiv:2107.12710

Tak H, Jung J-w, Patino J, Todisco M, Evans N (2021) Graph attention networks for anti-spoofing. arXiv. https://doi.org/10.48550/ARXIV.2104.03654. arXiv:2104.03654

Tan CB, Hijazi MHA, Khamis N, Zainol Z, Coenen F, Gani A et al (2021) A survey on presentation attack detection for automatic speaker verification systems: state-of-the-art, taxonomy, issues and future direction. Multimed Tools Appl 80(21):32725–32762

Tapkir PA, Patil HA (2018) Significance of teager energy operator phase for replay spoof detection. In: 2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), pp. 1951–1956. IEEE

Tapkir PA, Patil AT, Shah N, Patil HA (2018) Novel spectral root cepstral features for replay spoof detection. In: 2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), pp. 1945–1950. IEEE

Teng Z, Fu Q, White J, Powell ME, Schmidt DC (2022) Sa-sasv: An end-to-end spoof-aggregated spoofing-aware speaker verification system. arXiv preprint arXiv:2203.06517

Todisco M, Delgado H, Evans NWD (2017) Constant q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Comput Speech Lang 45:516–535

Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, Yamagishi J, Evans N, Kinnunen T, Lee KA (2019) Asvspoof 2019: future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441

Valero X, Alias F (2012) Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. IEEE Trans Multimedia 14(6):1684–1689

Wang Q, Lin X, Zhou M, Chen Y, Wang C, Li Q, Luo X (2019) Voicepop: a pop noise based anti-spoofing system for voice authentication on smartphones. In: IEEE INFOCOM 2019—IEEE conference on computer communications, pp. 2062–2070. https://doi.org/10.1109/INFOCOM.2019.8737422

Wang X, Yamagishi J, Todisco M, Delgado H, Nautsch A, Evans N, Sahidullah M, Vestman V, Kinnunen T, Lee KA et al (2020) Asvspoof 2019: a large-scale public database of synthesized, converted and replayed speech. Comput Speech Lang 64:101114

Witkowski M, Kacprzak S, Zelasko P, Kowalczyk K, Galka J (2017) Audio replay attack detection using high-frequency features. In Interspeech, pp. 27–31

Wong E, Sridharan S (2001) Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In: Proceedings of 2001 international symposium on intelligent multimedia, video and speech processing. ISIMP 2001 (IEEE Cat. No. 01EX489), pp. 95–98. IEEE

Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and counter measures for speaker verification: a survey. Speech Commun 66:130–153

Wu Z, Khodabakhsh A, Demiroglu C, Yamagishi J, Saito D, Toda T, King S (2015) Sas: A speaker verification spoofing database containing diverse attacks. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4440–4444. IEEE

Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilçi C, Sahidullah M, Sizov A (2015) Asvspoof 2015: the first automatic speaker verification spoofing and counter measures challenge. In: Sixteenth annual conference of the international speech communication association

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 counter measures challenge. IEEE J Sel Top Signal Process 11(4):588–604. https://doi.org/10.1109/JSTSP.2017.2671435

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 counter measures challenge. IEEE J Sel Top Signal Process 11(4):588–604

Wu X, He R, Sun Z, Tan T (2018) A light cnn for deep face representation with noisy labels. IEEE Trans Inf Forensics Secur 13(11):2884–2896

Wu Z, Das RK, Yang J, Li H (2020) Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. arXiv. https://doi.org/10.48550/ARXIV.2009.09637. arXiv:2009.09637

Wu H, Liu S, Meng H, Lee H-y (2020a) Defense against adversarial attacks on spoofing counter measures of ASV

Wu H, Liu AT, Lee H-y (2020b) Defense for black-box attacks on anti-spoofing models by self-Supervised learning

Xie Y, Li Z, Shi C, Liu J, Chen Y, Yuan B (2021) Enabling fast and universal audio adversarial attack using generative model. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp. 14129–14137

Xue J, Fan C, Yi J, Wang C, Wen Z, Zhang D, Lv Z (2023) Learning from yourself: a self-distillation method for fake speech detection. In: ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096837

Yang J, Das RK (2019) Low frequency frame-wise normalization over constant-q transform for playback speech detection. Digit Signal Process. https://doi.org/10.1016/j.dsp.2019.02.018

Yang J, Das R (2019) Long-term high frequency features for synthetic speech detection. Digit Signal Process 97:102622. https://doi.org/10.1016/j.dsp.2019.102622

Yang J, Das RK, Zhou N (2019) Extraction of octave spectra information for spoofing attack detection. IEEE/ACM Trans Audio Speech Lang Process 27(12):2373–2384. https://doi.org/10.1109/TASLP.2019.2946897

Zhang Y, Jiang F, Duan Z (2021) One-class learning towards synthetic voice spoofing detection. IEEE Signal Process Lett 28:937–941. https://doi.org/10.1109/LSP.2021.3076358

Zhang Y, Zhu G, Duan Z (2022) A probabilistic fusion framework for spoofing aware speaker verification. arXiv preprint arXiv:2202.05253

Zhou X, Garcia-Romero D, Duraiswami R, Espy-Wilson C, Shamma S (2011) Linear versus mel frequency cepstral coefficients for speaker recognition. In: 2011 IEEE workshop on automatic speech recognition & understanding, pp. 559–564. IEEE

Authors and Affiliations

Awais Khan[1] · Khalid Mahmood Malik[1] · James Ryan[1] · Mikul Saravanan[1]

✉ Khalid Mahmood Malik
mahmood@oakland.edu

Awais Khan
awaiskhan@oakland.edu

James Ryan
jaryan3@oakland.edu

Mikul Saravanan
mikulsaravanan@gmail.com

[1] Department of Computer Science and Engineering, Oakland University, Rochester 48309, MI, USA