Uncovering the Deceptions: An Analysis on Audio Spoofing Detection and Future Prospects

Rishabh Ranjan, **Mayank Vatsa**, **Richa Singh**
Indian Institute of Technology, Jodhpur, India
{ranjan.4, mvatsa, richa}@iitj.ac.in

Abstract

Audio has become an increasingly crucial biometric modality due to its ability to provide an intuitive way for humans to interact with machines. It is currently being used for a range of applications including person authentication to banking to virtual assistants. Research has shown that these systems are also susceptible to spoofing and attacks. Therefore, protecting audio processing systems against fraudulent activities such as identity theft, financial fraud, and spreading misinformation, is of paramount importance. This paper reviews the current state-of-the-art techniques for detecting audio spoofing and discusses the current challenges along with open research problems. The paper further highlights the importance of considering the ethical and privacy implications of audio spoofing detection systems. Lastly, the work aims to accentuate the need for building more robust and generalizable methods, the integration of automatic speaker verification and countermeasure systems, and better evaluation protocols.

1 Introduction

Voice as a biometric modality has been used for both identification and verification tasks, and it has a wide range of real-world applications such as banking, government, and law enforcement operations. For instance, Citi[1] uses voice samples to authenticate persons using automatic speaker verification (ASV) systems. Voice-enabled devices and virtual assistants such as Google Home and Microsoft Cortana are commonly used at home to help us manage our daily tasks. The worldwide voice recognition market has surpassed $3.5 billion in 2021 and is expected to reach $10 billion by 2028[2]. With the host of applications, there come several challenges associated with it. These voice-enabled devices often store a large amount of personal information and speech samples, and this data can be used to imitate one's voice. In 2020, an employee cloned the voice of the company's CEO and committed a financial fraud of $35 million. With the advancements in deep learning, it has become very easy to clone someone's voice (e.g., it has been claimed that Microsoft's VALL-E[3] can clone someone's voice in 3 seconds).

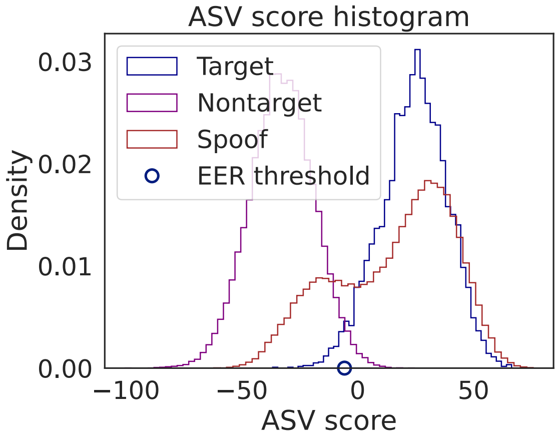

Figure 1: The score distribution of target users, non-target users and spoofed imposters. Due to an overlap in the distribution between the target and the spoofed imposter, models misclassify between bonafide and spoofed samples. The scores are taken from the ASVSpoof2019 dataset.

ASV systems [Kinnunen and Li, 2010] authenticate a speaker's identity by analysing their voice's unique characteristics. These systems work by analysing the characteristics of a person's voice, such as pitch, tone, and accent, to determine if they are who they claim to be. One of the main challenges with using voice as an identity verification method is the possibility of spoofing attacks [Evans *et al.*, 2013; Evans *et al.*, 2014; Ranjan *et al.*, 2022]. The speaker verification scores from the ASVspoof2019 dataset, as shown in Figure 1, illustrate that the score distribution of the spoofed imposter is similar to the target user. This suggests that spoofed samples can be easily authenticated as the target users.

Audio spoofing attacks are broadly classified into two categories: physical access attacks and logical access attacks. Physical access attacks are those that are introduced at the microphone (source) level, such as replay attacks and impersonation attacks. The different kinds of attacks are shown in

[1] https://tinyurl.com/citibankvoice

[2] https://tinyurl.com/voicemarket

[3] https://tinyurl.com/vallemicro

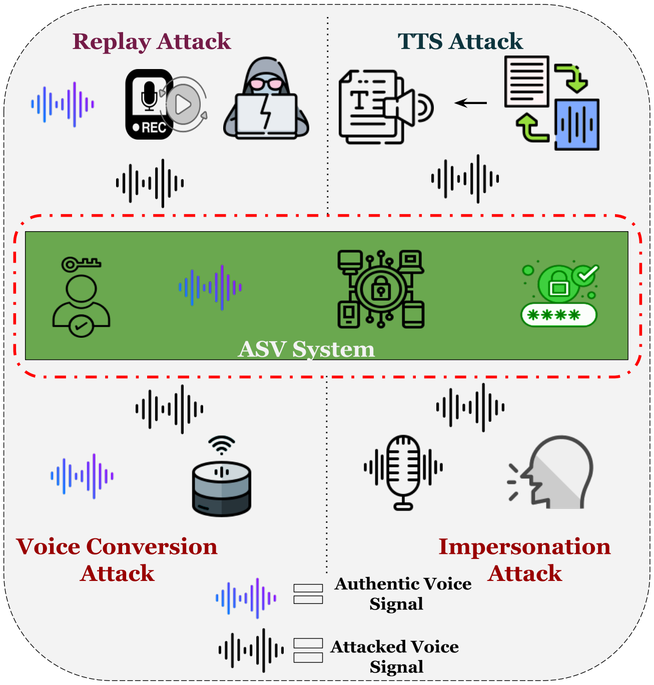

Figure 2: Illustration of different kinds of attacks on the ASV System. Replay and Impersonation attacks can be considered Physical Access attacks, while TTS and VC attacks are considered Logical Access attacks.

Figure 2. In Replay attacks, an attacker records a legitimate user's voice and then plays it back to the ASV system to gain access. This can be done by recording the user's voice during a legitimate login or by intercepting and recording the user's voice during a phone call. Logical access attacks, on the other hand, are introduced at the transmission level. These types of attacks include text-to-speech attacks and voice-conversion attacks. In text-to-speech attacks, an attacker uses a computer program to generate a voice that mimics a legitimate user's voice. This can be done by providing the text input to the program and generate speech that sounds like the legitimate user's voice. In voice-conversion attacks, the attacker uses a legitimate user's voice to generate an artificial voice to match the characteristics of the targeted speaker's voice. Recent studies [Kreuk *et al.*, 2018] have shown that ASV systems are also vulnerable to adversarial attacks. Several countermeasures are proposed to detect audio spoofing attacks. Most existing countermeasures are based on detecting artifacts in the generated speech. These artifacts can also be called a model's signature. The countermeasure systems take speech waveform as input and classify the input speech into bonafide or spoof.

To provide a more comprehensive understanding, several survey papers [Wu *et al.*, 2015a; Tan *et al.*, 2021] are published on audio spoofing detection. These papers have examined different aspects of spoofing attacks and the countermeasures that have been developed to combat them. The motivation for this research arises from the ongoing discussions and debates about the most effective methods and techniques for building countermeasure systems. This survey aims to provide a comprehensive overview of the current state-of-the-art in building a countermeasure system and identify the gaps or challenges that still need to be addressed. We summarize the main contributions as follows:

- This paper presents an analysis of speech generation and associated detection techniques,
- We provide a summary of evaluation metrics and existing datasets along with their limitations, and
- We present a meta-analysis and discuss the open challenges and future prospects in audio spoofing detection.

2 Evaluation Metrics

This section briefly introduces several standard performance metrics to understand better how to assess spoofing detection and speaker recognition systems.

Equal error rate (EER): EER is a commonly used performance metric for evaluating audio spoofing detection systems. The miss and false alarm rates of the spoofing detection system are increasing and decreasing with respect to different thresholds. The point at which the miss rate and false alarm rate become equal is called an Equal Error Rate (EER).

Detection Error Tradeoff (DET) curve: DET curve is a graphical representation of the performance of an audio spoofing detection system. It plots the false acceptance rate (FAR) on the x-axis and the false rejection rate (FRR) on the y-axis. The DET curve is similar to the Receiver Operating Characteristic (ROC) curve, but it is specifically used for detection tasks where the costs of false positives and false negatives are different.

Tandem Decision Cost function: The better-performing countermeasure system is not guaranteed to have a more reliable speaker verification performance. However, the CM system's performance directly impacts the ASV system's reliability. Thus, it is required to evaluate the performance of CM and ASV systems jointly. A new metric t-DCF [Kinnunen *et al.*, 2018] is proposed, which could bridge the gap between the CM system and the ASV system and can also be used to evaluate the performance of the CM system in isolation from the ASV system. The detection cost function is based upon the costs of missing the target users and falsely accepting the imposter users. The proposed t-DCF metric also considers the spoofing imposters and associated costs with accepting or rejecting them. The CM system and ASV system can be integrated in three ways. The metric considers four kinds of costs: a) the cost of the ASV system rejecting a target trial, b) the cost of the ASV system accepting a nontarget trial, c) the cost of CM rejecting a human trial, d) the cost of CM accepting a spoof trial.

3 Spoofed Speech Generation Techniques

There are several algorithms to create spoofed speech samples depending on the kind of attack. Recent advancements in deep learning have led to the development of robust and sophisticated models for fake audio generation, such

Dataset	Language	Attack type	No of Speakers	No of Samples
YOHO [Kreuk *et al.*, 2018]	English	Mimicry	2	960
WSJ [Ergünay and others, 2015]	English	SS and VC	283	-
SAS [Wu *et al.*, 2015b]	English	SS and VC	Real: 106 and Fake: 106	2,12,000
ASVspoof 2015 [Wu *et al.*, 2015c]	English	SS and VC	Real: 106 and Fake: 106	16651 Real + 255904 Fake
ASV Spoof Noisy Database	English	SS and VC	Real: 106 and Fake: 106	16651 Real + 255904 Fake
RedDots [Kinnunen *et al.*, 2017b]	5 languages	Replay	89	3750
ASVspoof 2017 [Kinnunen *et al.*, 2017a]	English	Replay	Real: 42 and Fake: 42	3566 Real + 14,466 Fake
ASVspoof 2019 [Wang *et al.*, 2020b]	English	SS,VC and Replay	Real: 106 and Fake: 106	121971 LA + 22157 PA
FOR [Reimao and Tzerpos, 2019]	English	SS	Real: 140 and Fake: 33	195000
ASVspoof2021 [Delgado *et al.*, 2021]	English	SS,VC, Replay and Deepfakes	Real:149 and Fake:149	-
HAD [Yi *et al.*, 2021]	Chinese	Human	Real: 218 and Fake: 218	-
WaveFake [Frank and Schönherr, 2021]	English, Japanese	TTS	Real: 2 and Fake: 2	117985

Table 1: Summary of existing audio spoof detection datasets.

as WaveNet [Oord *et al.*, 2016], and Tacotron [Wang *et al.*, 2017], capable of generating high-quality speech that is difficult to distinguish from real speech. In the following subsections, we discuss the different attack generation technologies.

3.1 Speech Synthesis

Speech synthesis is a technology that generates speech from a given input, such as text or speech. The process of generating speech from text is called text-to-speech synthesis, and changing the characteristics of a speaker's voice to make it sound like another speaker is called voice conversion. A typical speech synthesis model has three stages, the Input Analysis phase, the Acoustic Model phase and the Vocoder phase. In the input analysis phase, the input features are converted into linguistic features such as phonemes; the acoustic model is responsible for converting linguistic features into acoustic features such as spectrogram. Then, the vocoder converts the spectrogram into audio signals. As technology has progressed, computer-based speech synthesis methods have evolved, starting with early methods such as articulatory synthesis [Coker, 1976], formant synthesis[Seeviour *et al.*, 1976], and concatenative synthesis [Olive, 1977], and moving on to more advanced methods such as statistical parametric speech synthesis (SPSS) [Yoshimura *et al.*, 1999] and neural network-based speech synthesis [Wang *et al.*, 2017]. These neural network-based methods use deep learning techniques to generate speech and have shown great promise in producing highly realistic and natural-sounding speech.

[Zen *et al.*, 2013] use SPSS to generate audio waveforms. This works by converting text into linguistic features and then using the DNN to generate acoustic features. This approach was better than the HMM-based speech synthesis model as it could model complex context dependencies. Instead of generating text using three different modules (analysis, acoustic and vocoder), Tacotron [Wang *et al.*, 2017] is an end-to-end RNN-based, text-to-speech-system which uses an encode-decoder-based attention network and uses Griffin-Lim algorithm to generate the raw waveforms. Due to the recurrent nature of RNN, the encoder-decoder framework cannot be trained in parallel. With advancements in CNNs, authors proposed Deepvoice3 [Ping *et al.*, 2018], a fully convolutional neural network that generated mel-spectrogram instead of complex linguistic features. The growing use of transformers motivated the researchers to propose transformerTTS [Li *et al.*, 2019], which is based on an encoder-decoder-based attention network that can generate mel-spectrograms from phenomes. GANs are also used for speech synthesis. Multi-SpectroGAN [Lee *et al.*, 2021b] can be used to train the multi-speaker model with only the adversarial feedback. This model uses a conditional discriminator and can synthesize the mel-spectrogram without using re-construction loss between ground truth and the generated mel-spectrogram. The existing speech synthesis detection dataset, such as ASVSpoof2019 LA [Wang *et al.*, 2020b] uses NN-based, SPSS-based TTS systems. The dataset uses WORLD [Kawahara, 2006] and STRAIGHT [Kawahara, 2006] as vocoders for generating the raw audio. The existing datasets are shown in Table 1. However, the existing dataset lacks samples from sophisticated TTS systems that use diffusion models [Lee *et al.*, 2021a].

3.2 Replay Attacks

In a Replay attack, the attacker intercepts the speech of the legitimate user, records it and plays it back later to impersonate the original speaker. This is the easiest kind of attack as it does not require any technical expertise. The existing Replay attack generation techniques primarily focus on single-hop attacks. Datasets like ASVspoof 2017 [Kinnunen *et al.*, 2017a] and ASVspoof 2019 [Wang *et al.*, 2020b] only have single-playback recordings (i.e. played once), so they cannot assess anti-spoofing algorithms against multi-playback attacks. Also, the existing datasets need to consider microphone array characteristics, which are crucial in modern VCDs. Datasets like Voice Spoofing Detection Corpus (VSDC) [Baumann *et al.*, 2021] handle this problem by proposing multi-hop replay attacks. The multi-hop replay attack problem can be considered a multi-class classification problem instead of the binary-class classification problem. This would assist in evaluating countermeasure systems in more realistic situations than in a lab environment. The existing Replay attack detection datasets are shown in Table 1.

3.3 Adversarial Attacks

The adversarial attacks are performed by adding small and imperceptible perturbations to the original audio data that can fool the system into producing incorrect predictions. This started with [Szegedy *et al.*, 2014] showing that adversarial examples can fool image classification models. This led to adversarial attacks becoming a topic of significant interest in image processing, leading to the development of numerous successful attack techniques, including the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015] and Basic Iterative Method (BIM) [Kurakin *et al.*, 2018].

Adversarial attacks in the audio field have gained growing attention from [Kreuk *et al.*, 2018], following the progress made in adversarial attacks for images. There are two main categories of adversarial attacks on Speaker Recognition systems: optimization-based attacks [Goodfellow *et al.*, 2015; Kreuk *et al.*, 2018] and signal processing-based [Wang *et al.*, 2020a] attacks. Optimization-based attacks generate adversarial examples by solving an optimization problem that formalizes the goal of the attack. This method typically involves minimizing a distance metric between the original and adversarial audio signals, subject to some constraints, such as the adversarial example being imperceptible to humans or having a small perturbation size. On the other hand, signal processing-based attacks use signal processing techniques to manipulate the original audio signal in a way that causes a desired misclassification by the SRS. These techniques can include adding noise, changing the frequency spectrum, or modifying the phase of the audio signal.

4 Spoofed Speech Detection Techniques

This section investigates different algorithms used to detect spoofing attacks in audio recordings. There are two main approaches to this problem, machine learning and deep learning-based algorithms. Most early spoof detection algorithms (before 2015) are based on traditional machine learning-based models (GMM, HMM, SVM). However, in recent years, deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have become increasingly popular for this task.

4.1 Traditional ML-based Approaches

[Satoh *et al.*, 2001] propose a synthetic speech detector to protect text-based ASV systems. The proposed model calculates the inter-frame difference of log-likelihood scores of the speaker's Gaussian Markov Model. The proposed architecture is only evaluated for the HMM-based speech synthesis model. This was the first-ever attempt to detect voice-based spoofing. This method was not evaluated for Replay attack detection. For replay attack detection, a similarity-based model [Shang and Stevenson, 2010] is proposed, which calculates the similarity between test occurrence and recordings of the speaker. The main drawback of the paper is that it needs N-recorded speech for attack detection. [Alegre *et al.*, 2013] use long-range features for synthetic speech detection. The authors extract utterance-level features extracted from frame-level features. The authors also show that certain ASV systems are more robust to spoofing attacks. The electromagnetic interference of audio recording creates an Electric Network Frequency (ENF) signal. [Su *et al.*, 2013] propose a decorrelation operation to extract ENF signal, which can be used for Replay attack detection.

Prior to 2015, most of the research on speech spoofing detection was conducted on private datasets, preventing the researchers from benchmarking the results. This led to the creation of ASVspoof2015 [Wu *et al.*, 2015c], which contains samples for TTS and VC attacks. [Todisco *et al.*, 2017] extracted Constant-Q Cepstral Coefficients (CQCCs) features based on constant Q transform (CQT) to detect spoofing attacks of the ASVSpoof2015 dataset. The authors have evaluated the proposed system on two existing datasets and the results have highlighted the sensitivity towards attacks. The authors did not evaluate their CQCC features on Replay attacks. Later, a new dataset ASVspoof2017 [Wu *et al.*, 2017] is prepared to spearhead the research in detecting Replay attacks.

[Font *et al.*, 2017] compared different handcrafted features such as Constant-Q Cepstral Coefficients (CQCCs) and Mel Frequency Cepstral Coefficients (MFCCs) to train a GMM-based model. The authors found that Subband Spectral Centroid Magnitude Coefficients (SCMCs) performed better than all other features. However, the model could not generalize when evaluated on the cross-corpus database. [Yang *et al.*, 2019] propose using power information of CQCC features using multi-level transform. The multi-level transform is applied to the power information of CQCC features.

In 2019, a more diverse dataset, ASVSpoof2019 [Wang *et al.*, 2020b], is released, containing synthetic attacks from 6 known and 13 unknown systems. The dataset also contains replayed attacks from 6 known systems. The researcher [Malik *et al.*, 2020] showed that replay attacks could be modelled as a non-linear process, creating harmonic distortion in the signal. These harmonic distortions can be exploited to detect the attacks. The authors also propose a multi-hop replay attack dataset VSDC. The replayed attack samples of the ASVspoof2019PA dataset are replayed once and are easier to detect than the multi-hop replay attacks. A fusion of different features like MFCC, GTCC, Spectral Flux, and Spectral Centroid is explored for audio spoofing detection. The proposed architecture is only evaluated for synthetic speech detection.

4.2 Deep Learning-based Approaches

Deep learning-based techniques are extensively used for spoofing detection. The deep-learning-based approaches are used both for feature extraction and as classifiers. [Chen *et al.*, 2015] have used deep learning for the audio spoofing detection task for the first time. This architecture uses deep learning architecture to extract learned representation. The authors compute a compact feature representation which uses distance metrics for classification. This algorithm extracts *s-vector* features from Deep Neural Network (DNN) and uses Mahalanobis distance for spoofing detection. Convolutional long short-term neural network (CLDNN) [Dinkel *et al.*, 2017] uses raw audio signal for spoof detection. The CLDNN uses layers that learn dependencies between time and frequency Domains. The architecture is evaluated on ASVspoof2015 and BTAS2016 datasets. However, the proposed architecture is used for both replay and speech synthesis attacks. Replay noise [Shim *et al.*, 2018] is explored by researchers for replay attack detection. The authors propose a Multi-Task learning framework to exploit this replay noise. However, the multi-task architecture is only evaluated for a single dataset.

Residual connection-based approaches are also explored for audio spoofing detection. [Alzantot *et al.*, 2019] have built three different variants of residual networks. The authors observe that the fusion of models outperforms individual models. The fusion model achieves zero EER and zero t-DCF on the development set of the ASVspoof2019LA dataset. The zero EER and t-DCF show that the model is overfitting on the development set. The results show that fusion of five networks performed better than single systems and baseline architectures. This architecture can detect both replay and speech synthesis attacks. Authors have yet to experiment with the unified model. A deep neural network (DNN) based pipeline [Lai *et al.*, 2019] is proposed for spoof detection. The proposed pipeline uses four stages feature engineering, DNN models, network optimization and system combination. The authors have explored multiple features and different variants of residual networks and squeeze-excitation networks. [Wu *et al.*, 2020] showed that countermeasure system against adversarial PGD attack and proposed spatial smoothing and adversarial training [Wu *et al.*, 2020] defence against an adversarial attack on countermeasure systems trained on ASVspoof2019 dataset.

Several countermeasures use handcrafted features extracted from a raw audio signal to train deep neural networks. Motivated by the vision transformers [Dosovitskiy *et al.*, 2021], [Zhang *et al.*, 2021b] use transformers and residual networks to detect fake speeches. The encoder of the transformer is used to extract deep features, which the residual network uses for classification. [Ranjan *et al.*, 2022] uses Spectral and Temporal features extracted from RawNet2 encoder for audio spoofing detection.

A capsule network with modified-routing algorithm [Luo *et al.*, 2021] is proposed for audio spoofing detection. The authors argue that capsules can learn better representations than convolutional networks. Motivated by residual connections, Several architectures [Tak *et al.*, 2021a] based on ResNet are proposed. The paper [Tak *et al.*, 2021b] proposes several modifications to RawNet2 architecture for audio spoofing detection. [Jung *et al.*, 2022a; Tak *et al.*, 2021a] have also used Graph Neural Networks (GNNs) for spoofing tasks. [Jung *et al.*, 2022a] propose a spectral and temporal graph attention layer for spoof detection. The architecture achieved state-of-the-art performance on the AVSpoof2019LA dataset. The authors have also proposed lightweight architecture for the deployment of edge devices. The architecture is not evaluated for Replay attacks. Anti-spoofing systems only detect spoofing, but a combined system can be developed by incorporating speaker verification. Spoofing-aware speaker verification (SASV) challenge [Jung *et al.*, 2022b] for the first time integrated speaker-verification and anti-spoofing. [Zhang *et al.*, 2022] propose a probabilistic fusion method for joint integration of ASV and CM systems. The authors use a product and fine-tuning strategy to get the SASV scores. Due to a limited dataset for evaluating the ASV and CM systems, the generalizability of the system could not be evaluated. Lightweight countermeasure [Liao *et al.*, 2022] is proposed for developing countermeasures systems for edge devices. Authors use Speaker distillation and generalized pre-training to train ResNetSE architecture effectively. With advancements in attack generation techniques, the future development of countermeasure systems lies in end-to-end deep learning-based countermeasure systems.

4.3 Meta-Analysis

In this section, we discuss the trends in developing countermeasure systems. We rank architectures with respect to their performance in detecting different attacks.

Logical Access Attacks: These included TTS and voice cloning attacks. The ASVSpoof2019LA dataset contains samples of logical access attacks. Figure 3 shows the performance of different architectures on the ASvspoof2019LA dataset. Four of the five best systems use raw waveforms as input to the architecture, and the remaining one [Zhang *et al.*, 2021a] takes spectrogram as input. This shows that raw audio waveforms have complementary information with respect to handcrafted features. All the top five performing architectures used residual connections in their architecture. All top 3 performing architectures are based on raw audio and Graph neural networks. AASIST-L [Jung *et al.*, 2022a] is a lightweight model with just 85K parameters and it is in the top 2 performing systems. CQCC features are the most informative features used by the authors in architecture MCG-Res2Net50 [Li *et al.*, 2021]. However, none of the architecture is evaluated on a cross-corpus database. This limits the generalizability of the architectures. The results suggest that raw audio waveforms and deep learning-based systems can be used to build a countermeasure system for logical access attack detection.

Physical Access Attacks: These generally contain samples of Replay attacks. The attack samples are collected in 27 different acoustic environments. Figure 4 shows the result for different architectures on the ASVSpoof2019PA dataset. The results suggest that hand-crafted features are essential for audio spoofing detection. The features extracted using CQT are extensively used for spoofing detection. Many architectures use ResNet and Light CNN as backbone classifiers. However, most systems are not evaluated on multiple attacks and cross-database settings.

5 Open Challenges

This section discusses the challenges and future prospects for research in ASV systems.

5.1 Diverse Dataset Collection

Several challenges are associated with collecting datasets that can be used to train the countermeasure system for spoof detection. The challenges arise while collecting real audio data and generating fake audio. We list the key challenges as follows:

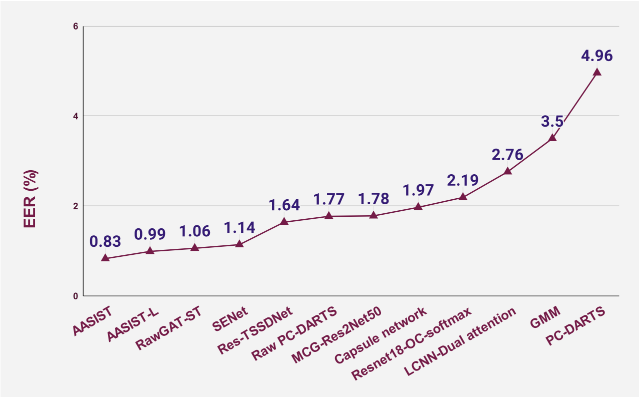

Figure 3: Performance of countermeasure system on the ASVSpoof2019LA dataset. The dataset contains samples of TTS and VC attacks.

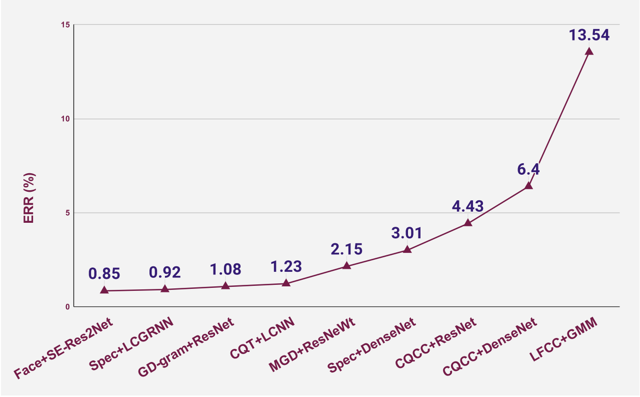

Figure 4: Performance of countermeasure system on the ASVSpoof2019PA dataset. The dataset contains samples of Replayed attacks.

Diversity of Dataset: To effectively detect spoofing attacks, it is crucial to have a diverse set of audio data that accurately reflects the target population. In the case of India, this would mean taking into account a wide range of demographic factors such as gender, age, dialect, and language. It is essential to have a balanced representation of different genders and age groups in the dataset, to ensure that the detection algorithms can accurately identify spoofing attacks regardless of the speaker's gender or age. Dialects and language also play a crucial role in Indian society, with a vast array of regional dialects and multiple official languages. Hence, a balanced representation of the population is also essential.

Quality of Data: The quality of the data and the labelling of that data are critical to the success of an audio spoof detection task. A high-quality dataset should have accurate and detailed annotations, including speaker identity, gender, age, and dialect. Additionally, the recording conditions should be standardized, with high-quality equipment and real-world environments. A model only trained on data collected in a controlled environment may not perform well in real-world scenarios due to differences in acoustic conditions, background noise, and other environmental factors.

Privacy Preservation: Privacy concerns are a significant challenge in collecting bonafide audio data for audio spoof detection. Audio data can contain personal information that needs to be protected, such as names, addresses, and sensitive information. Data collection must comply with legal constraints for data protection, such as the European General Data Protection Regulation (GDPR). Privacy regulations aim to protect personal information from unauthorized access and misuse, and researchers must ensure that the data collection process adheres to these regulations.

Accent-Aware Synthetic Data: Synthetic data should reflect the language and accent of the target speaker to be impersonated, making cross-lingual compatibility an essential factor to consider in creating synthetic data. The synthetic data should also reflect those specific accents if the target speaker speaks English with a different accent, such as an Indian or Chinese accent. This requires a comprehensive understanding of different accents and their characteristics and using language-specific tools and techniques to create synthetic data that accurately reflects each accent. If the synthetic data does not reflect the target speaker's accent, it may lead to incorrect results in the detection algorithms or biases in the performance of the algorithms.

Multi-Speaker Dataset: One of the biggest challenges while collecting synthetic data for low-resource languages like Hindi is the limited availability of high-quality speech data. In many cases, the amount of data available for low-resource languages is significantly smaller than that for high-resource languages, which can lead to overfitting and poor generalization performance. So, to perform better, synthetic data should reflect the language and accent of the target speaker to be impersonated, making cross-lingual compatibility an essential factor to consider in creating synthetic data.

5.2 Robust Spoof Detection Algorithms

Most of the existing detection algorithms need to be more generalizable. Spoof detection models are evaluated on limited and non-diverse data, leading to an inaccurate assessment of their generalization ability. Existing spoof detection models outperform benchmark datasets but fail in real-world scenarios. Also, the existing evaluation metrics must adequately capture the models' generalization ability.

5.3 Cross-Corpus Evaluation

Most existing spoof detection models are not evaluated on cross-corpus, which is a crucial step in developing audio spoof detection models. It helps to determine their generalizability and robustness to variations in recording conditions, data distribution, and noise types. Such evaluations are necessary because a model trained on a specific corpus may perform well on that particular dataset but not on others due to different data distribution and recording conditions. [Korshunov and Marcel, 2016] show that the existing spoofing detection systems need to be more generalizable as they perform poorly when tested on the same attack samples from a different dataset. The data variability can lead to significant differences in model performance across different corpora, making it essential to evaluate models against diverse datasets.

5.4 Language Independent Models

People use speaker verification systems in their native languages to authenticate themselves. However, existing countermeasure systems are heavily language dependent. This poses a severe risk to the ASV systems. This possible solution is that separate models must be trained and deployed for each language, which can be time-consuming and computationally intensive, increasing the complexity of building and deploying audio spoof detection systems. Developing language-independent models that can effectively detect audio spoofing across multiple languages would greatly benefit the field and increase the efficiency of audio spoof detection systems.

5.5 Universal Spoof Detection Model

There exist a few models that can detect all types of spoofing attacks (replay, TTS, VC, adversarial attacks). This is because different types of attacks require different methods of detection. For example, detecting replay attacks might require examining the properties of a microphone, whereas detecting adversarial examples might require identifying changes in the statistical properties of the data. Therefore, creating a universal model for all types of attacks is difficult. The universal model can detect a broader range of attacks, improving security in the field and making attacks less likely to go unnoticed. Additionally, deploying a single model costs less than deploying multiple models. This is because it reduces the need for multiple systems and personnel to manage and maintain each model. However, a general-purpose model may not perform as well as a specialized model for a particular type of attack, as it requires a trade-off between accuracy and performance.

5.6 Privacy Preserving Spoof Detection Models

As data privacy becomes increasingly essential, Federated Learning (FL) provides a solution that allows organizations to collaborate and share information while preserving the privacy of individual data. This is particularly relevant in tasks where biometric information, such as speech, is collected. FL enables multiple parties to collaborate to train a neural network without exchanging data samples, ensuring individual data privacy. However, there are various challenges while training a federated learning-based spoof detection model. One of the significant challenges is non-IID data distribution, where each participant has their own data distribution, which may differ from other participants' data distribution. This results in a non-IID data scenario and can lead to less effective learning compared to centralized training methods.

5.7 Interpretable and Explainable Models

Explainability is very important in developing spoofed detection models, as it allows us to better understand these models' decision-making processes. This is important for several reasons. First, explainability helps ensure that models are not biased toward specific classes or features that can negatively impact performance. It then provides insight into how the model processes the input data so that further improvements can be made. Finally, it gives a better understanding of model limitations, and helps identify and fix potential errors. For example, in the image domain, CNNs have given excellent results for tasks such as image classification. However, the decision-making process of these models still needs to be improved to interpret. To address this, researchers proposed using explainable AI (XAI) techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping)[Selvaraju *et al.*, 2020] which provides a visual explanation of CNN prediction. This gives a better understanding of the features that the model uses to make predictions. Along the same lines, [Ge *et al.*, 2022] use SHapley Additive exPlanations (SHAP) to describe a detection model for audio spoofing. However, this method only works for confirmed attacks.

5.8 Fair Spoof Detection

Fairness and bias are crucial in developing spoof detection models as they directly affect the performance and accuracy of the models. Bias in these models can result in unequal treatment of different groups, leading to unfair and unreliable decisions. To ensure that these models are fair, it is essential to consider demographic information and various factors that may impact the accuracy of the model's predictions. Techniques such as debiasing the training data and using gender-neutral and diverse datasets can be employed to address this. However, there is no dataset to evaluate the fairness of the existing countermeasure system. The research community needs a dataset that has labels about the gender, ethnicity, and accent of the speakers.

6 Conclusion

The advancements in attack generation methods have motivated researchers to develop countermeasures systems for audio spoofing detection. In this paper, we review the existing literature for audio spoofing detection. Despite the advancements in audio spoofing detection methods, several research gaps exist. We discuss the research gaps and open challenges that require proper attention and continuous research efforts. Solving or providing solutions for these open research problems is necessary for building trusted audio spoofing detection systems.

Acknowledgements

This research is supported by a grant from MHA, India. M. Vatsa is also partially supported through the SwarnaJayanti Fellowship by the Government of India.

References

[Alegre *et al.*, 2013] F Alegre, A Amehraye, et al. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *ICASSP*, pages 3068–3072, 2013.

[Alzantot *et al.*, 2019] M Alzantot, Z Wang, et al. Deep residual neural networks for audio spoofing detection. In *Interspeech*, pages 1078–1082, 2019.

[Baumann *et al.*, 2021] R Baumann, K M Malik, et al. Voice spoofing detection corpus for single and multi-order audio replays. *Comput. Speech Lang*, 65:101132, 2021.

[Chen *et al.*, 2015] N Chen, Y Qian, et al. Robust deep feature for spoofing detection — the SJTU system for ASVspoof 2015 challenge. In *Interspeech*, pages 2097–2101, 2015.

[Coker, 1976] C H Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, 1976.

[Delgado *et al.*, 2021] H Delgado, N Evans, et al. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535*, 2021.

[Dinkel *et al.*, 2017] H. Dinkel, N. Chen, et al. End-to-end spoofing detection with raw waveform cldnns. In *ICASSP*, pages 4860–4864, 2017.

[Dosovitskiy *et al.*, 2021] A Dosovitskiy, L Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Ergünay and others, 2015] S K Ergünay et al. On the vulnerability of speaker verification to realistic voice spoofing. In *BTAS*, pages 1–6, 2015.

[Evans *et al.*, 2013] N Evans, T Kinnunen, and J Yamagishi. Spoofing and countermeasures for automatic speaker verification. In *Interspeech*, pages 925–929, 2013.

[Evans *et al.*, 2014] N W. D. Evans, T Kinnunen, et al. Speaker recognition anti-spoofing. In *Handbook of Biometric Anti-Spoofing - Trusted Biometrics under Spoofing Attacks*, pages 125–146. Springer, 2014.

[Font *et al.*, 2017] R Font, J M Espín, et al. Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge. In *Interspeech*, pages 7–11, 2017.

[Frank and Schönherr, 2021] J Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. In *NeurIPS Datasets and Benchmarks*, 2021.

[Ge *et al.*, 2022] W Ge, M Todisco, et al. Explainable Deepfake and Spoofing Detection: An Attack Analysis Using SHapley Additive exPlanations. In *Odyssey*, pages 70–76, 2022.

[Goodfellow *et al.*, 2015] I J. Goodfellow, J Shlens, et al. Explaining and harnessing adversarial examples. In *ICLR (Poster)*, 2015.

[Jung *et al.*, 2022a] J-w Jung, H-S Heo, et al. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP*, pages 6367–6371, 2022.

[Jung *et al.*, 2022b] J-w Jung, H Tak, et al. SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. In *Interspeech*, pages 2893–2897, 2022.

[Kawahara, 2006] H Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27:349–353, 2006.

[Kinnunen and Li, 2010] T Kinnunen and H Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[Kinnunen *et al.*, 2017a] T Kinnunen, Md Sahidullah, et al. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Interspeech*, pages 2–6. ISCA, 2017.

[Kinnunen *et al.*, 2017b] T H. Kinnunen, Md. Sahidullah, et al. Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. *ICASSP*, pages 5395–5399, 2017.

[Kinnunen *et al.*, 2018] T Kinnunen, K-A Lee, et al. t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. In *Odyssey*, pages 312–319, 2018.

[Korshunov and Marcel, 2016] P Korshunov and S Marcel. Cross-Database Evaluation of Audio-Based Spoofing Detection Systems. In *Interspeech*, pages 1705–1709, 2016.

[Kreuk *et al.*, 2018] F Kreuk, Y Adi, et al. Fooling end-to-end speaker verification with adversarial examples. In *ICASSP*, pages 1962–1966, 2018.

[Kurakin *et al.*, 2018] A Kurakin, I J. Goodfellow, and S Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112, 2018.

[Lai *et al.*, 2019] C-I Lai, N Chen, et al. ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks. In *Interspeech*, pages 1013–1017, 2019.

[Lee *et al.*, 2021a] S-G Lee, H Kim, et al. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2021.

[Lee *et al.*, 2021b] S-H Lee, H-W Yoon, et al. Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis. In *AAAI*, pages 13198–13206, 2021.

[Li *et al.*, 2019] N Li, S Liu, et al. Neural speech synthesis with transformer network. In *AAAI*, pages 6706–6713, 2019.

[Li *et al.*, 2021] X Li, X Wu, et al. Channel-wise gated res2net: Towards robust detection of synthetic speech attacks. In *Interspeech*, pages 4314–4318, 2021.

[Liao *et al.*, 2022] Y-L Liao, X Chen, et al. Adversarial speaker distillation for countermeasure model on automatic speaker verification. In *SPSC*, pages 30–34, 2022.

[Luo *et al.*, 2021] A Luo, E Li, et al. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP*, pages 6359–6363, 2021.

[Malik *et al.*, 2020] K. M. Malik, A. Javed, et al. A light-weight replay detection framework for voice controlled iot devices. *JSTSP*, 14(5):982–996, 2020.

[Olive, 1977] J Olive. Rule synthesis of speech from dyadic units. In *ICASSP*, pages 568–570, 1977.

[Oord *et al.*, 2016] A v d Oord, S Dieleman, et al. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016.

[Ping *et al.*, 2018] W Ping, K Peng, et al. Deep voice 3: 2000-speaker neural text-to-speech. *ICLR*, 2018.

[Ranjan *et al.*, 2022] R Ranjan, M Vatsa, et al. Statnet: Spectral and temporal features based multi-task network for audio spoofing detection. In *IJCB*, pages 1–9, 2022.

[Reimao and Tzerpos, 2019] R Reimao and V Tzerpos. For: A dataset for synthetic speech detection. In *SpeD*, pages 1–10, 2019.

[Satoh *et al.*, 2001] T Satoh, T Masuko, et al. A robust speaker verification system against imposture using an HMM-based speech synthesis system. In *Interspeech*, pages 759–762, 2001.

[Seeviour *et al.*, 1976] P Seeviour, J Holmes, and M Judd. Automatic generation of control signals for a parallel formant speech synthesizer. In *ICASSP*, pages 4860–4864, 1976.

[Selvaraju *et al.*, 2020] R R. Selvaraju, M Cogswell, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020.

[Shang and Stevenson, 2010] W Shang and M Stevenson. Score normalization in playback attack detection. In *ICASSP*, pages 1678–1681, 2010.

[Shim *et al.*, 2018] H-J Shim, J-W Jung, et al. Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In *TAAI*, pages 172–176, 2018.

[Su *et al.*, 2013] H Su, R Garg, et al. Enf analysis on recaptured audio recordings. In *ICASSP*, pages 3018–3022, 2013.

[Szegedy *et al.*, 2014] C Szegedy, W Zaremba, and other. Intriguing properties of neural networks. In *ICLR*, 2014.

[Tak *et al.*, 2021a] H Tak, J-w Jung, and other. Graph attention networks for anti-spoofing. In *Interspeech*, pages 2356–2360, 2021.

[Tak *et al.*, 2021b] H Tak, J Patino, et al. End-to-end anti-spoofing with rawnet2. In *ICASSP*, pages 6369–6373, 2021.

[Tan *et al.*, 2021] C B Tan, Hijazi, et al. A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. *Multim. Tools Appl.*, 80(21):32725–32762, 2021.

[Todisco *et al.*, 2017] M Todisco, H Delgado, et al. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.*, 45:516–535, 2017.

[Wang *et al.*, 2017] Y Wang, R. J. S-Ryan, et al. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*, pages 4006–4010, 2017.

[Wang *et al.*, 2020a] Q Wang, P Guo, and L Xie. Inaudible adversarial perturbations for targeted attack in speaker recognition. In *Interspeech*, pages 4228–4232, 2020.

[Wang *et al.*, 2020b] X Wang, J Yamagishi, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.*, 64:101114, 2020.

[Wu *et al.*, 2015a] Z Wu, N W. D. Evans, et al. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.*, 66:130–153, 2015.

[Wu *et al.*, 2015b] Z Wu, A Khodabakhsh, et al. SAS: A speaker verification spoofing database containing diverse attacks. In *ICASSP*, pages 4440–4444, 2015.

[Wu *et al.*, 2015c] Z Wu, T Kinnunen, et al. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech*, pages 2037–2041, 2015.

[Wu *et al.*, 2017] Z Wu, J Yamagishi, et al. Asvspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE JSTSP*, 11(4):588–604, 2017.

[Wu *et al.*, 2020] H Wu, S Liu, et al. Defense against adversarial attacks on spoofing countermeasures of ASV. In *ICASSP*, pages 6564–6568, 2020.

[Yang *et al.*, 2019] J Yang, R K Das, et al. Extraction of octave spectra information for spoofing attack detection. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(12):2373–2384, 2019.

[Yi *et al.*, 2021] J Yi, Ye Bai, et al. Half-truth: A partially fake audio detection dataset. In *Interspeech*, pages 1654–1658, 2021.

[Yoshimura *et al.*, 1999] T Yoshimura, K Tokuda, et al. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Eurospeech*, pages 2347–2350, 1999.

[Zen *et al.*, 2013] H Zen, A Senior, and M Schuster. Statistical parametric speech synthesis using deep neural networks. In *ICASSP*, pages 7962–7966, 2013.

[Zhang *et al.*, 2021a] Y Zhang, W Wang, et al. The effect of silence and dual-band fusion in anti-spoofing system. In *Interspeech*, pages 4279–4283, 2021.

[Zhang *et al.*, 2021b] Z Zhang, X Yi, et al. Fake speech detection using residual network with transformer encoder. In *IH&MMSec Workshop*, pages 13–22, 2021.

[Zhang *et al.*, 2022] Y Zhang, Ge Zhu, et al. A probabilistic fusion framework for spoofing aware speaker verification. In *Odyssey*, pages 77–84, 2022.