International Journal of Basic and Applied Sciences, 14 (5) (2025) 224-235

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/w36qrh81
Research paper

Advancements in Voice Spoofing Detection: A Comprehensive Review

Rekha Rani [1] *, Bal Kishan [2]

[1] *Research Scholar, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India*
[2] *Associate Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India*
**Corresponding author E-mail: rekha.rs.dcsa@mdurohtak.ac.in*

Received: July 13, 2025, Accepted: August 27, 2025, Published: September 7, 2025

Abstract

The reliability of voice-based authentication has increased with the adoption of voice-controlled technologies and digital transactions. Automatic Speaker Verification (ASV) provides a dependable approach due to its special capacity to confirm identity based on speech. ASV is mostly used in telecommunications, banking, law enforcement, and smart assistants to increase security and user comfort. However, spoofing attacks like voice conversion and speech synthesis are increasingly targeting these systems, making them less compatible, examining responses to new kinds of attacks through data augmentation, and highlighting the role of transfer learning in improving detection even when there is a lack of data. This review discusses the importance of strengthening ASV systems with data augmentation to address new threats, transfer learning to enhance detection with limited data, and adaptive models to keep up with advancing spoofing attacks.

Keywords: *ASV; Spoofing Attacks; Adversarial Attack; Machine Learning; Countermeasures.*

1. Introduction

Authentication is critical for information security, allowing only authorized users to have access to the data. While commonly used biometric techniques such as fingerprints and iris scans, are common, voice-based authentication provides a touchless, convenient option.

Automatic Speaker Verification (ASV) systems identify distinctive vocal traits to verify one's identity and are widely used in various applications like banking, law enforcement, virtual assistants, and secure communications. These systems are vulnerable to spoofing attacks, where malicious actors manipulate or generate voice samples to bypass security. With the increasing use of AI-based voice synthesis and conversion techniques [1], spotting authentic voices from imitations is getting harder and harder. Most common features like background noise, channel variations, and speaker aging are affecting the performance of ASV systems, leading to the creation of more complex ASV technologies that can manage real-world challenges effectively.

1.1. Background and Motivation

This review highlights recent advancements in ASV security, such as adversarial training to strengthen system robustness, self-supervised learning to improve performance with limited data, and data augmentation to expand flexibility. To ensure that user data are not manipulated, methods that preserve privacy are also examined [4]. With these modifications, we highlight the growing need for secure and reliable ASV systems and suggest strategies to strengthen voice-based authentication against possible spoofing attacks. Figure 1 also shows that the total amount of recorded spoofing occurrences has slowly climbed up over the past few years. This makes it more crucial than ever for organizations to come up with strong defences right away.

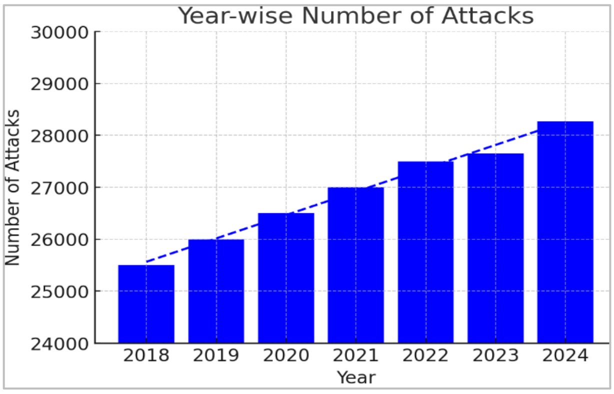

Fig. 1: Evolution of Voice Attacks Over the Years. Information Gathered from Online Databases (Google Search, Industry Surveys) and Openly Available Reports. This Number Might Not Be A Precise Estimate of All Worldwide Statistics, But It Does Show A Rising Tendency.

ASV systems are widely used in biometric authentication, secure communication, voice-controlled virtual assistants, and forensic analysis [3]. It is widely used because of its simplicity and hands-free control. As presented in Figure 1 [2], the number of spoofing attacks has increased from 25,500 in 2018 to nearly 28,300 in 2024 due to advancements in voice synthesis and voice conversion technologies used by attackers.

Table 1 illustrates ASV in comparison to other biometric technologies, highlighting that although it is touchless and device-independent, it remains more vulnerable to spoofing attempts. Unlike fingerprint or iris recognition systems, ASV systems must be robust against background noise, speaker aging, and spoofing attacks to maintain reliability [5]. Foolproof security measures are essential to ensure the effectiveness of ASV in real-world applications.

Table 1: A Brief Comparison Analysis of ASV With Other Biometric Modalities, Highlighting Relative Vulnerabilities to Spoofing, Environmental Sensitivity, and Cost Factors (Compiled from Biometric Studies

Attributes	ASV (Voice)	Fingerprints	Iris Recognition	Face Recognition
Collected Data	Speaking voice	Fingerprint image	Iris image	Image of facial expression
User Functionality	Touchless	Contact required	Position required	Good light/brightness required
Safety/Security	Moderate (very sensitive to attack)	High (difficult to forget)	High (unique pattern)	Moderate (can be attacked by AI)
Individuality (unrepeatable)	Moderate (varies according to age/health)	High (unique pattern)	High (unique pattern)	Moderate (varies according to age)
Estimated Cost	Moderate (according to the used software)	Low (inexpensive)	High (special scanner required)	Moderate (depending on camera cost)
Robustness (dependability)	Can be used on various devices	Require scanner	Require specialized cameras	Require cameras
Noise sensitivity	Medium (background noise)	Low (not affected)	Low (not affected)	Moderate (light affection)

2. Voice Spoofing Attacks

Voice spoofing is a serious threat to voice verification systems, allowing attackers to bypass security using techniques such as voice morphing, speech synthesis, and impersonation [11]. Through mimicking real voices, attackers can acquire unauthorized permission to sensitive information like personal data and financial accounts [12]. Voice spoofing attacks are divided into two categories—direct spoofing and indirect spoofing—as illustrated in Fig. 2. These classifications highlight emerging adversarial threats that take advantage of model deficiencies as well as capturing conventional spoofing techniques.

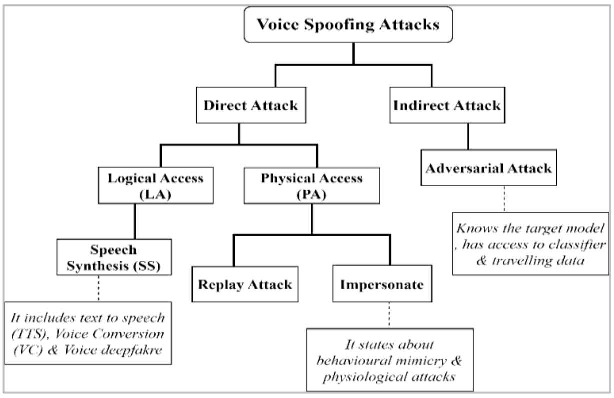
Voice Spoofing Attacks
Direct Attack
Indirect Attack
Logical Access (LA)
Physical Access (PA)
Adversarial Attack
Knows the target model , has access to classifier & travelling data
Speech Synthesis (SS)
Replay Attack
Impersonate
It includes text to speech (TTS), Voice Conversion (VC) & Voice deepfakre
It states about behavioural mimicry & physiological attacks

Fig. 2: Organizes Speech Spoofing Assaults into Two Categories: Direct (Physical/Logical Access) and Indirect (Adversarial).

Voice spoofing attacks refer to tactics used to fool voice authentication systems. There are mainly two types of these attacks: Direct Attacks and Indirect Attacks.

- Direct Attacks include both Logical and physical access attacks, in which attackers convert their voices into exact copies. Four primary categories of voice spoofing exist:

1) Speech Synthesis: This technique creates a synthetic voice from text using a text-to-speech (TTS) system to generate commands, requests, and passwords [7].
2) Voice conversion: In this technique, the attacker mimics the target's voice using their natural voice.

3) Replay Attack: This involves recording a speaker's voice and then playing it back as if it were being uttered in real time.
4) Impersonation: In this technique, a human speaker mimics the voice of a different individual [8].

Indirect Attacks or Adversarial Attacks [3] alter voice features during processing, making detection more difficult. These attacks, which are particularly tricky for machine learning-based systems, make use of model defects rather than altering the raw audio as direct attacks do. Knowledge about various adversarial attack methods and their impacts is essential to voice authentication security enhancements [17]. Nowadays, adversarial attacks are becoming more common, as attackers alter voice characteristics to fool detection systems. Table 2 below categorizes different adversarial attack techniques, stating how methods like combined noise, spectrogram variations, and feature manipulation fool models by manipulating important attributes. These attacks confuse systems trying to differentiate between authentic and synthetic voices.

To defeat these emerging challenges, defense techniques such as noise injection, adversarial training, and multimodal detection (combining audible and visual signals like lip movements) are being applied [12]. As adversarial attacks become more robust, improving detection models to secure voice authentication systems offers a challenge.

Table 2: Common Adversarial Attack Techniques in Voice Spoofing Detection, with Descriptions, Their Impacts on ML/DL Models, and Defense Strategies

Adversarial Attack Methods	Descriptions	Effects on Detection Models	Security Capability
Combined Noise	Modifying an audio clip of a fake voice by adding background noise, white noise, or other aberrations	This could cause the detection model's features to malfunction, misclassifying anything as real speech	During training, use noise injection to enhance data
Variations in Spectra	Adjusting a fake voice sample's spectrogram by slightly altering its frequency or amplitude	Can modify the model's interpretation of vocal traits, possibly avoiding detection	Using spectrogram-based adversarial training
Disturbance in voice conversion [15]	Creating a fake voice by making minor adjustments to the voice conversion method	It may fool the model into believing that the faked voice is real because of its changed features.	Using strong voice conversion methods that are harder to manipulate
Manipulated Features	Directly altering audio properties (such as MFCCs) to produce hostile examples	Can evade detection if a certain feature, such as Mel-frequency cepstral coefficients (MFCCs)—a representation of the short-term power spectrum of sound—is significantly relied upon by the model	Methods of feature normalization to lessen sensitivity to minute changes
Adversaries with Morphed Voices [8]	Fusing elements of a fake and real voice to produce a hybrid hostile	Utilizes traits from both real and counterfeit voices, possibly complicating the detection model.	Multimodal spoofing detection that takes into account extra data (such as lip movements)

The continuing development of spoofing beyond audio manipulation is highlighted by these adversarial schemes, and they call for robustness-ensuring countermeasures, including adversarial training, feature normalization, and multimodal fusion.

3. Literature Review

Research communities have conducted several surveys on automatic speaker verification (ASV) systems and their detection methods. The current survey primarily focuses on system spoofing attacks.

The researchers present an overview of various technologies and advancements in automatic speaker verification systems. It highlights how crucial having a defense against spoofing attacks is. It also suggests developing hybrid systems, which integrate many technologies.

Most recent work focuses on deep learning models, like CNNs, WaveNet, and RNNs, for binary classification and restricted audio features. It investigates multiclass and binary classification and contrasts isolated methods with no performance comparison [2]. Huang Yida et al. state better ways to spot fake voices. Convolutional neural networks (CNNs) were mostly used in traditional methods, but have issues with long-range relationships. Attention-based architectures, especially Transformers, have shown promising results over long distances but need a lot of time in training. However, the prime drawback is that its performance frequently differs among datasets (for example, ASVspoof vs. VCC), highlighting that models well-suited for one circumstance may not adapt well to other situations. For good speech embedding extraction, a new study has shown how important it is to use both low-level and high-level features together. Present feature pyramid networks (FPNs) have some good points, but they have trouble with feature fusion, which means they might lose useful information when they mix multi-scale features. The AFP-Conformer, a model that blends CNNs and Transformers, is what Huang et al. say can be used to solve these issues. Using an asymptotic feature pyramid network method, this model makes it easier to collect both local and global data and lessens the amount of data that is lost during feature fusion. The tests they did showed big changes in how well things are found, especially when it comes to tricky spoofing attacks [6].

Prior research looks at restricted auditory characteristics in binary and multi-class classification. Since it solely employs lone algorithms, a performance comparison is not present on large datasets. In machine learning, the Clever Hans effect has been researched extensively. In training data, confounding variables often improve performance [7]. A robust detector for resource-constrained IoT devices, utilizing a "feature pyramid" to examine varied voice features and an "adaptive weighting" module to home in on crucial information [8]. A novel technique that uses acoustic characteristics to protect speaker verification against fraudulent impersonations. Spoof detection by a unique method that combines particular characteristics, conventional classification, and score-level fusion [9] can enhance cross-database identification with a 95% success rate in detecting synthetic sounds. Attackers create audio samples that sound like the victim. When a new joint loss function is developed, modern authentication schemes are successfully attacked with high success rates. It leaves voice authentication systems open to malicious party attacks [10]. The security of upcoming voice authentication systems against complex spoofing efforts is encouraged by these advancements. To manage the constantly changing environment, classifying methods, “incremental learning" strategy, and countermeasure approaches can protect voice authentication in the future [11].

Improving anti-spoof detection capabilities is the main objective, rather than classifying synthesis techniques. When qualities are classified, the detection of spoofing techniques becomes more reliable. By using the recommended multitask attribute classification training technique, the system's resilience will be improved [13].

El-Sayed Atlam et al. analyze 286 studies on deepfake (DF) spread on social media published from 2018 to June 2024. The study highlights significant gaps in the literature, such as the lack of studies on digital interventions (only 1%) and philosophical perspectives (3%). The findings emphasize the need for collaboration among experts in various fields, including computer vision, machine learning, and media forensics. Since machine learning fails to achieve robustness without considering social, ethical, and adversary factors, this multidisciplinary approach is necessary for ASV spoofing detection. 15].

In one study, CFCCIF-ESA, a newcomer, is compared with variables such as vector size, filter selections, and static versus dynamic properties [20] . The importance of feature selection in creating strong Automatic Speaker Verification (ASV) systems is highlighted by this focus, which opens the door to more precise and dependable spoof detection.

A wide range of datasets is being prepared to assess anti-spoofing techniques in stressful conditions, including the Voice Spoofing Detection Corpus (VSDC) [21]. One study focused on liveness identification and recorded an extensive range of 66 people's artificial and real voices, containing both remastered and actual recordings. Anti-spoofing options must be verified against several kinds of potential threats to guarantee reliable and robust performance [22].

In the hunt for complete spoofing detection, the hunt is still on. Diverse datasets, such as the Voice Spoofing Detection Corpus (VSDC) [21], are being constructed to assess anti-spoofing methods in challenging settings. An extensive collection of 66 people's genuine and synthetic voices, comprising both real-world and re-recorded recordings, was captured in one study that concentrated on liveness detection. It is crucial to test anti-spoofing systems against a variety of possible threats to ensure strong and dependable performance [22]. Nirupam Shome et al. (2023) focus on advancements in speaker recognition, particularly through deep learning techniques, highlighting the evolution of methods from traditional approaches to modern deep neural networks, which have significantly improved the performance. Various deep learning architectures, such as Stacked Restricted Boltzmann Machines (RBMs) and Encoder-Decoder networks built with Recurrent Neural Networks (RNNs), enhance better accuracy in recognizing speakers through feature extraction and classifications.

Barhoush, Mahdi, et al. The article "Speaker identification and localization using shuffled MFCC features and deep learning" talks about big steps forward in automatic speaker identification (ASI) and localization. It focuses on the role of microphone array processing in many areas, like videoconferencing and surveillance. The authors talk about the problems with ASI, especially when it comes to getting unique features from speech in different environments. It states that Mel Frequency Cepstral Coefficients (MFCC) are often used, but they don't always work well because of things like noise and reverberation.

The writers suggest a new method that uses Shuffled MFCC (SHMFCC) and its version, Difference Shuffled MFCC (DSHMFCC) to make the system more accurate in both single- and multi-speaker situations. The method works well even with small training datasets. This demonstrates a growing trend of lightweight yet effective models, an aspect still underexplored in voice spoofing detection. [23].

Mohd Rafiz Hanifa et al. Over the last eighty years, there has been a lot of change in the literature on speaker recognition. This is because technology has improved, and people want safer identification systems. Speaker recognition is a part of speech processing that tries to identify people by their unique voice traits, which can be affected by things like race, age, gender, and mood. The paper explains the difference between speaker recognition and speech recognition and stresses how important it is to know the difference between these terms to avoid confusion in study and use. Smart gadgets and voice assistants have made speaker recognition technology more common in everyday life, which has led to new developments in the field.

Additionally, the study touches on adversarial attacks that are capable of fooling machine learning algorithms, highlighting how crucial it is to have resilient frameworks to withstand such errors. This is consistent with recent ASVspoof trials, indicating that even the most advanced systems suffer significant performance declines in the presence of hidden attacks. [27].

Table 3: Depicts Major Challenges in Voice Spoofing Detection (2020-2024), Such as the Absence of A Universal Countermeasure and Cross-Dataset Generalization Problems. in Response to These, Incorporating Explainable AI with Data Augmentation Is Proposed

State-of-the-art	Year	Limitations in current studies
Hao Meng et al. [40]	2025	Lack of exploration of the range of attack types.
M A Basit et al. [24]	2024	Lack of global countermeasures against spoofing attacks.
Mahdi Barhoush et al. [33]	2024	Struggle with noise and reverberation.
Boyd Jason et al. [2]	2023	Lack of diverse model comparison and performance evaluation A small exploration of audio features leads to incomplete model decisions.
Zhou Jincheng et al. [19]	2022	Lack of robust countermeasures against voice replay attacks (noise, pitch) Feature extraction is also limited.
Dawood Hussain et al. [16]	2022	Detecting replay attacks for voice spoofing still needs improvement. The existing method lacks performance variability.
X. Wang et al. [38]	2021	Difficulty in generalization and classifier limitations Limited available datasets Lack of cross-dataset performance evaluations
R. Rahmeni et al. [15]	2020	Limitation in spoof detection performance on a specific database A challenge in securing voice biometric systems against spoofing attacks.
Our Contribution		Need for Explainable AI Integration and Data Augmentation Strategies

Our Contribution: The outcome of this review identifies novel gaps, such as the need for data augmentation strategies and Explainable AI integration, as well as generalized countermeasures.

Data Augmentation Strategies: The process of applying different modifications to existing data to create new training datasets, thereby improving model robustness and generalization. Common strategies include noise injection, pitch shifting, time stretching, and speed perturbation [37].

Explainable AI (XAI) integration: This involves developing techniques aimed at making machine learning models and their decisions understandable to humans. XAI techniques help to interpret why the model made a particular decision, such as speaker verification.

Generalized Countermeasures: Designing systems that offer reliable defence against a variety of spoofing scenarios while avoiding overfitting to a single dataset or attack type (more on this in Section 5).

Additionally, the development of generalized security authorities for identifying various spoofing techniques is discussed in Section 5 of this review.

4. Techniques Used in Spoofed Voice Detection

Voice spoofing, in which a person impersonates another person, is identified by machine learning algorithms. There are several methods, like acoustic analysis of audio recordings and deep vocal modeling, that can be used to achieve it. More accurate detection of spoofed voices is possible only because machine learning algorithms can identify patterns and characteristics by training models on vast datasets of real and spoofed voices.

Machine learning algorithms control feature learning and pattern classification by using important features for training to get it ready for testing. The test sample can be a real speaker, or it could be a known or unknown attacker.

There are two categories of speaker models: generative and discriminative. Hidden Markov Models (HMM) [4], i-vectors, and Gaussian Mixture Models (GMM) [8, 9, 30] are examples of generative models. Some of the distinctive models involve support vector machines (SVM) [22, 23, 24], deep learning [26], and neural networks [14]. In speaker verification, using a DNN improves accuracy when comparing different speakers. This paper describes techniques crucial to tackling new spoofing attacks as they balance the accuracy, robustness, and real-time detection capabilities required of authentic voice systems.

Regular models, such as GMMs and HMMs, execute well on small, controlled datasets, but they frequently fall short against contemporary TTS and VC systems due to their struggles with cross-dataset generalization and invisible attacks. On the other hand, richer temporal and spectral connections are captured by deep learning techniques (CNNs, RNNs, Transformers), but at the expense of increased processing demands. Current research revolves around this trade-off between robustness and efficiency, which is a significant deployment obstacle in actual ASV systems.

4.1. Approach Based on Machine Learning

While classifiers are a type of conventional machine learning method, generative methods represent another type. These methods can be utilized for spoof detection in datasets pertinent to the initial studies on ASV systems.

1) Gaussian Mixture Models (GMMs): GMMs serve as first-stage feature extraction tools utilized in voice spoofing detection. They can recreate the statistical distribution of features derived from speech, including Mel-Frequency Cepstral Coefficients (MFCCs). It helps to recognize variations from the usual distribution of genuine voice characteristics [8]. Although instead of spoofing cues, GMMs may overfit to dataset-specific artifacts due to the "Clever Hans" effect, which reduces their robustness across datasets like ASVspoof and VCC.
2) Hidden Markov Model (HMMs): HMM represents a widely-used approach for tasks involving voice recognition, speaker identification, and verification. These models are often used when designing TD-ASV systems. A statistical model [29] describes unobservable states that change over time. HMMs work well for sequential modeling, but they perform poorly against the highly convincing synthetic voices produced by existing deepfake systems.
3) Support Vector Machine (SVM): SVM addresses both regression and classification tasks. It serves as a classification model for voice spoofing detection, comparing authentic and spoofed voices. It serves by identifying the optimal decision boundary, based on parameters such as pitch, energy, and spectral features, to distinguish between genuine and fabricated voice data points [33]. Although SVMs are not scalable for high-dimensional spectrogram data, where deep learning is the dominant technique, they will typically perform far better than GMMs on just a handful of features.
4) K-means algorithm: For Speaker Verification, which is a classification task, unsupervised clustering serves as an appropriate technique. The K-means clustering algorithm can distinguish different clusters within the voice feature input vectors. The clustering method initializes k clusters with a minimum distortion between the vector and the centroid, then randomly assigns centroids to each cluster. By distributing the vectors, the algorithm can iteratively capture the cluster's minimum distortion value.

K-means provides an efficient unsupervised approach by clustering voice feature vectors into distinct classes. In speaker verification, this enables the model to identify anomalies potentially associated with spoofing attempts.

For efficient spoof determination, aggregation is not enough because many spoof samples have a significant feature set matching of real ones, prompting the use of supervised or hybrid techniques.

Real-World Consequences and Difficulties with Machine Learning: While these methods demonstrate strong detection capabilities, translating them into real-world systems presents several challenges. Machine Learning provides strong capabilities for voice spoofing recognition, but translating these developments into practical systems presents challenges and technical hurdles.

The price of computing is an important limitation. CNN and Transformer-based models, which obtain state-of-the-art EER reductions on ASVspoof 2019, consume a lot of GPU resources, which prohibits them from being implemented on mobile devices or the Internet of Things without improvement. Research shows that clipping methods and lightweight front-end designs are promising approaches for real-time innovative deployment (e.g., Wickramasinghe et al., 2023). This demonstrates that computational performance is both a technical and a deployment barrier, since models with high EER on ASVspoof benchmarks sometimes perform poorly in IoT or wireless settings because of resource constraints and latency.

Practical Consequences:

- Increased Security: Unauthorized access and fraudulent activity are reduced by identifying and stopping spoofing voices [37].
- Better User Experience: Voice authentication systems identify users and stop unwanted access attempts so users can feel secure using them.
- Decreased Expenses: Businesses and organizations can experience a reduction in expenses by implementing strong spoof detection to mitigate fraudulent activities and illegal access.

Latest trends in machine learning for spoofed voice detection:

- End-to-End Learning: Reducing the requirement for pre-defined features, this method trains a single model to extract features and perform classification (real/spoof) straight from raw audio input [35]. This simplifies the pipeline by reducing dependency on feature engineering, potentially allowing models to adapt better to new spoofing methods.
- Explainable AI (XAI): Recognizing the decision-making process of machine learning (ML) models becomes essential as they grow in complexity. Developers can enhance the interpretability of the model for spoof detection and uncover potential biases or flaws by utilizing XAI approaches [38]. XAI helps developers understand model decisions, ensuring that detection is not only accurate but also interpretable, which is vital for debugging and model improvement.
- Lightweight Real-Time Models: Machine learning models must be effective when deployed on devices with limited resources. For real-time spoof detection on embedded systems or mobile devices, research on lightweight architectures and effective training techniques is essential [35]. These models are critical for deploying efficient spoof detection on low-power devices like mobile phones, where real-time processing is necessary.

To minimize the size of models without dramatically sacrificing accuracy, some new attempts use knowledge distillation and quantization. It enables models to execute on-device while retaining an EER of less than 10% on ASVspoof benchmarks.

GMMs and SVMs are two machine learning techniques that have historically been advantageous in identifying simple spoofing patterns, nevertheless they frequently perform worse on heavier, more intricate datasets. Researchers are turning to deep learning techniques, which utilize the benefits of deeper speech signal representations and provide better generalization across invisible threats, to get beyond these restrictions.

4.2. Deep learning-based approach

These techniques are more suitable for the audio spoofing community since they perform better on large datasets with intricate distribution structures. These methods can be effectively used with both raw speech signals and feature vectors produced by various feature extraction techniques.

1) Deep Neural Network (DNN): It acts as a feature for end-to-end speaker verification. During the training time, the loss function, representation level, and speaker model all have an impact on the speaker representation [26]. DNNs excel at mapping complex, non-linear relationships within voice data, enhancing feature discrimination. Considering that DNNs are excellent at mapping intricate, non-linear relationships in voice data, require a lot of processing power and massive databases with labels. Considering this, they are fairly accurate but less suitable for deployment on devices with limited resources.
2) Recurrent Neural Network (RNN): The temporal history of the voice stream can be recorded by a recurrent neural network. Because RNN is a member of the sequence-based scoring group, it does not modify the network in any way during the training or assessment stages [29]. RNNs, due to their sequence-based structure, capture temporal patterns in voice data, making them effective for analyzing continuous voice. Researchers opt for novel modifications like LSTM or GRU for spoof detection on account of their slower training times as well as frequent vanishing gradient problems.
3) Convolutional Neural Network (CNN): CNN is primarily used to find observable, local patterns in the dataset that help differentiate between authentic speech. It takes less preparation of the data for these networks because they use kernels on convolutional layers. The pooling layer is the fundamental structural element of the CNN. It reduces the specific size of the dataset to facilitate computations [32]. CNNs identify spatial patterns in spectrograms, allowing them to detect nuanced spectral features between real and spoofed voices. While CNNs are delivering state-of-the-art results in spoof detection, they may not be able to handle newly identified spoofing schemes and take up huge training datasets to prevent overfitting.

These models work together to capture valuable parts of voice, such as spectral patterns (CNNs), abstract non-linear features (DNNs), and temporal dynamics (RNNs/LSTMs), strengthening robustness against various spoofing techniques.

Latest trends in deep learning for spoofed voice detection:

- Self-Supervised Learning: Methods like Wav2Vec 2.0 and HuBERT train models using massive unlabeled datasets. Even though these models outperform trained techniques in generalizing to invisible spoofing assaults, they are still restricted due to their high training costs (GPU resources, time) [36].
- Multi-modal Learning: Combining audio with complementary modalities (e.g., lip movements, facial expressions) strengthens spoof detection. In video-audio pairings, this cross-modal consistency check is especially useful to spot deepfakes.
- Deployment on Edge Devices: Research is moving toward lightweight architectures (e.g., MobileNet, TinyCNN, quantized RNNs) to enable real-time on-device detection [40]. An ongoing problem is striking a balance between accuracy and reduced processing footprint, particularly for IoT and mobile banking applications.

Practical Challenges and Emerging Solutions: Machine learning and deep learning methods offer powerful tools for detecting voice spoofing. However, real-world applications require fast processing, cross-dataset generalization, and interpretability. Challenges include:

- Adaptation to Novel Attacks: Models trained on ASVspoof datasets may underperform on real-world spoofing attacks (dataset bias).
- Computational Constraints: Large models (CNNs, Transformers) demand extensive GPU resources, limiting feasibility on smartphones.
- User Trust and Transparency: Black-box DL models may be accurate but lack interpretability.

Federated learning for privacy-preserving training, model compression (quantization, pruning) for resource efficiency, and Explainable AI (XAI) to boost model prediction probability are some instances of these Emerging Methods.

Table 4 shows different methods for detecting spoofed voices, categorized into feature extraction, detection models, and fusion approaches. Methods for extracting features like MFCC and i-vectors help in deciding necessary voice characteristics, while deep speaker embeddings enhance real-time detection. Detection models like GMMs, CNNs, and LSTMs evaluate these features to distinguish between authentic and faked voices, with deep learning models giving enhanced accuracy. Score-level fusion, a type of fusion approach, combines multiple methods to improve performance. This review helps in understanding how different methods perform and their pros and cons, and the common performance metrics (such as EER, t-DCF, and Accuracy) used to evaluate them.

Table 4: Comparison of Voice Spoofing Detection Techniques

Category	Methods	Description	Strength	Weakness	Performance Metrics
Feature Extraction Techniques	MFCC (Mel-Frequency Cepstral Coffecient)	Extracts characteristics from speech signals with an emphasis on spectral data.	Widely used and simple to implement.	Limited capacity to detect sophisticated spoofing methods (like deepfakes).	EER, Accuracy
	i-vector & x-vector	Methods for extracting features using factor analysis.	More resilient to fluctuations and noise than MFCC.	Costly to compute.	EER, t-DCF
	Deep speaker embedding techniques	Learn comparative representations that are distinct to each speaker.	Effective in real-time scenarios.	Prone to adversarial attacks intended to deceive the embedding model	EER, Detection Accuracy
Classification & Detection Models	Gaussian Mixture Models (GMMs) [32]	Types of statistics that depict real and fake voice distributions.	Effective at capturing overall voice characteristics.	Tending to overfit training data.	EER, Accuracy
	Convolutional Neural Network (CNN) [32]	Develop sophisticated voice data representations to differentiate real from fake.	Excellent accuracy combined with deepfake detection powers.	Big datasets are necessary for training.	EER, t-DCF, Precision/Recall

	Long Short-Term Memory (LSTMs)	Captures temporal dependencies in speech signals.	Effective for detecting temporal inconsistencies.	Computationally expensive.	EER, Accuracy, F1 Score
	Transformer-Based Models (e.g., Wav2Vec, HuBERT)	Self-supervised learning models for robust feature extraction.	Handles large-scale voice data effectively.	Requires extensive training and resources. High resource demand, not ideal for edge devices.	EER, t-DCF, AUC (Area Under Curve)
Fusion & Decision Strategies	Score-level fusion	Combines the results of various detection techniques.	Enhances overall performance by utilizing the advantages of several techniques.	Increases complexity and necessitates fine-tuning certain techniques.	EER, t-DCF, Fusion Cost Metrics

5. Contributing Datasets and Evaluation Measures

Every day, many scientists generate and utilize fresh datasets because researchers require a diverse range of data to work with. Every dataset possesses unique characteristics of its own. Besides the data, the methods researchers use also affect the results of their studies. Some datasets do exceptionally well when it comes to false-positive rates, while others perform better in other domains. Therefore, determining which dataset or technique yields the best results is not necessary [39].

5.1. Dataset Used

In the early research on voice spoofing detection, several voice and speaker identification databases were used; however, after 2015, the scientific community released several evaluation databases, such as the RedDots, ASVspoof 2015, ASVspoof 2017, and ASVspoof 2019 challenge databases. The ASVspoof database offers replay spoofing attacks in addition to Speech Synthesis and Voice Conversion spoofing techniques. It was designed to imitate Physical Attacks and Logical Attacks [31].

ASVspoof 2015: For both LA and PA attacks, speech synthesis and voice conversion are used to create real and fake voices in the database. There is no background noise or channel effects. Training, Development, and Evaluation are the three components that make up the database [20].

Focus: The main objective of this dataset was to identify attacks that impersonate voice conversion (VC) and speech synthesis (TTS).

Data: It included 106 speaker recordings, each of which included both real and fake samples produced by different TTS and VC algorithms.

Challenges: The dataset is smaller than later versions, which presents a challenge. It also have limited applicability to most recent or advanced attacks because it concentrated mostly on particular spoofing techniques.

ASVspoof 2017: Because the RedDots corpus and its replayed speech served as the foundation for the ASVspoof 2017 Challenge database, it is a replay database with text-dependent speech [19], [27].

Focus: The purpose of this dataset is to identify replay attacks, which use spoofing to playback captured real voice.

Data: It contained 210 speaker recordings of real speech samples as well as replayed versions that had been altered using different replay settings (such as noise addition or channel distortion).

Challenges: The dataset size is still small in comparison to new benchmarks, and replay assaults might be less common than synthetic-based spoofing methods.

ASVspoof 2019: The main performance statistic for ASVspoof 2019 is the Tandem Detection Cost Function (t-DCF), which was recently proposed, along with (%) EER [10] [22] [31].

Focus: It includes both replay and synthesis spoofing attempts.

Data: Compared to earlier iterations, it included recordings from more than 1000 speakers and various spoofing strategies and modifications.

Challenges: Larger data sets enhance precision but need high processing capacity to process and analyze.

Vox Celeb: It is a collection of audio-visual material that has been taken from YouTube videos. The speakers in the data set have a nice mix of nationalities; there are speakers with American, Finnish, Indian, and other accents. Of the speakers, 39% are women and 61% are men.

Focus: VoxCeleb is a massive dataset of celebrity video interviews that was initially not intended for spoofing detection.

Data: Thousands of celebrity recordings with authentic and varied speech samples are included. Spoof detection models, especially those that use speech synthesis to simulate human voices, can be trained using this data.

Challenges: VoxCeleb does not include spoof samples; it needs to generate synthetic copies or utilize anomaly detection techniques to spot any spoofing attempts in real data.

RedDots: These datasets have a greater number of recording sessions with fewer spoken English words in each. Providing 52 sessions every week to each speaker for a year was the aim of creating the dataset. 24 sentences total—10 common, 10 unique, 2 free choices, and 2 free texts—were covered in these two-minute sessions. The inter- and intra-speaker types in this sample vary greatly.

Focus: Like ASVspoof 2017, the focus of this dataset is on replaying assaults. It is more difficult to record in real-world settings with fluctuations and background noise.

Data: Replayed versions of RedDots have over 100 speakers' recordings from a variety of settings and situations, altered with distortions and real-world background noise.

Challenges: The real-world nature of the data makes analysis more difficult and may present new difficulties for spoof detection programs. Furthermore, generalizability may be limited because the dataset size may be smaller than that of ASVspoof 2019.

An overview of several methods for speaker verification and spoofing detection across several datasets is given in Table 5. Each study's features, classifiers, and datasets are highlighted. On the VoxCeleb dataset, for example, Aakshi Mittal used cepstral coefficients with machine learning and deep learning models; on the ASVspoof 2015 dataset, Jason Beyond et al. used spectral, cepstral, and pitch-related features with CNN, WaveNet, GRU, and LSTM classifiers. The variety of approaches, features, and classifiers used on various datasets is displayed in the table, which represents the changing field of voice authentication research. The rapid development of voice spoofing analysis is visible in this comparison, as datasets are gradually growing and realistic qualities are being captured to capture a variety of assault situations.

Table 5: Highlights How Datasets Described in Section 5.1 (E.G., ASVspoof, Voxceleb, Reddots) Have Been Widely Used Across Studies with Different Feature Sets and Classifiers. This Comparison Underlines Not Only the Diversity of Available Resources but Al

Authors	Methods		Datasets
	Features	Classifiers	
Aakshi Mittal [1]	Cepstral coefficient	ML, DL Models	VoxCeleb1 & VoxCeleb2
Jason Beyond et al. [2]	Spectral, cepstral, and pitch related features	CNN, WaveNet, GRU, LSTM	ASVspoof 2015
Muhammad Sajjad et al. [3]	Spectral Centroid	CNN, LSTM (Long Short-Term Memory), GRUs	CASIA, ATVS-FFp, LivDet 2013
Andre Kassis [4]	LPMS, MFCC	GMM, CNN	ASVspoof 2019
Ali Javed et al. [5]	ATCoP, GTCC	SVM	VSDC, ASVspoof 2019, Google's LJ Speech
Yequin Ren et al. [8]	FPM, FAM	GMM	ASVspoof 2019
Joan Monterio et al. [9]	Deep bottleneck feature and frame-level descriptors	GMM, CNN, ResNet, PLDA (Probabilistic Linear Discriminant Analysis) used for binary classification	ASVspoof 2019
Jincheng Zhou et al. [10]	GCC, MFCC	Bi-directional LSTM	ASVspoof 2019 PA dataset
Ankur T Patil et al. [11]	TEO, CFCCIF-ESA	GMM, CNN	ASVspoof 2015
Hao Zhou et al. [12]	MFCC, LPC (Linear Predictive Coding)	SE- DenseNet, SE-Res2Net	ASVspoof 2019, LA, PA Dataset
Jinlin Guo et al. [14]	MFCC	ML, DL	ASVspoof 2019
Raoudha Rahmeni et al. [15]	MFCC, SCC, LogFBE	XGBoost tree boosting Algorithm	ASVSpoof 2015
Hussain Dawood et al. [16]	CLS-LBP (Center- Symmetric Local Binary Pattern)	LSTM	ASVspoof 2019
Raoudha Rahmeni et al. [17]	IAIF	SVM, ELM	ASVspoof 2019
Huixin Liang et al. [18]	STFT	CNN	ASVspoof 2019
Joao Monterio et al. [19]	2D CNN	Ensemble-based approach	ASVspoof 2017
Ivan Himawan et al. [20]	FC2 layer	CNN, DET (detection error tradeoff)	BTAS 2016 and ASVspoof 2015
Junxiao Xue et al. [22]	MFCC, LPC, LPCC (Linear Predictive Coding Coefficients), LFCC	CNN, SVM	ASVspoof 2019
Awais Khan et al. [23]	Cepstral coefficient, Linear predictor coefficient	GMM, SVM, CNN and CNN- GRU	VSDC datasets, ASVspoof 2019, 2021, ReMASC, RedDots, Vox Celeb, etc.
Muhammad Abdul Basit et al. [24]	2D Convolution and ProdSpec features	SVM, Naïve Bayes	ASVspoof 2019
Souvik Sinha et al. [26]	SSL	DNN	ASVspoof 2019
Buddhi Wickramasinghe et al. [27]	MFCC	Bi-directional LSTM	ASVspoof 2017
Ariel Cohen et al. [30]	Double-sided log spectrogram (DSL)	GMM	ASVspoof 2021, LA, DeepFake

As seen in Table 5, the choice of dataset, features, and classifiers significantly influences detection performance, which makes evaluation metrics (discussed in Section 5.2) equally critical for fair benchmarking.

5.2. Evaluation measures

ASV systems are assessed using metrics to quantify their effectiveness and pinpoint areas for growth, just like athletes use metrics to monitor their progress. False rejections (letting in imposters) and false acceptances (denying genuine) users of the system are considered by all evaluation measures. These metrics help compare spoofing detection methods fairly and identify areas of improvement.

Equal Error Rate (EER): The EER is the point at which the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. The overlapping area's EER indicates the proportion of instances in the system reacts either by admitting a fraudulent speaker or rejecting a legitimate one [37].

Formula: EER= FAR

EER False Acceptance Rate and FAR False Acceptance Rate

False Acceptance Rate (FAR): This is the proportion of instances in which a system misidentifies a speaker who is impersonating someone else as authentic. It is preferable to have a smaller FAR since it shows that imposters can mislead the system less easily [32].

Formula: FAR = FA /TA

FA = No. of false acceptance and TA = No. of total attempts (voices)

False Rejection Rate (FRR): Indicates the proportion of instances in which the system wrongly dismisses a legitimate speaker. It is preferable to have a lower FRR as it shows that access denials are less likely to occur [34].

Formula: FRR = FR/ TA

FR = No. of false rejection and TA = No. of total attempts (voices)

Detection Error Tradeoff (DET) Curve: Graphical plot of FAR vs. FRR at various thresholds. A lower and steeper DET curve suggests greater performance since it signifies a reduced trade-off between FAR and FRR [32].
Area Under the ROC Curve (AUC): The performance of a model at various categorization thresholds is shown graphically by the Receiver Operating Characteristic (ROC) curve. Plotting the True Positive Rate (TPR) on the y-axis and the FAR on the x-axis. The total area under the ROC curve is denoted by AUC [36].

Higher AUC = better discrimination

Tandem Decision Cost Function (t-DCF): (Most Recent, Sophisticated Metric) that examines the ASV system and its defensive capabilities simultaneously. It is particularly valuable for security-critical applications since it considers the cost of various error scenarios [37].

t-DCF = Cn * FAR * Cs + Cm * FRR * Ct is the formula. [37]

Cn: The cost of accepting a fake acceptance (a voice impersonation)
Cm: The price of turning away a legitimate user due to a mistaken rejection.
Ct: The price of a target trial for actual users
Cs: The price of a fraudulent trial.
Note: This was the primary metric in ASVspoof 2019, emphasizing security-focused evaluation [36].

Table 6: Summary of Evaluation Metrics in Voice Spoofing Detection

Metrics	What it Measures	Why to Use
EER	Where FAR=FRR	Single-value comparison
FAR	% of accepted spoofed voice	Lower FAR value = strong detection
FRR	% of rejected real voice	Lower FRR value = More user-friendly
DET Curve	Trade-off between FAR & FRR	Visual check of performance
AUC	Area under the ROC curve	Higher AUC = better detection of real vs. fake
T-DCF	Countermeasure cost + ASV	Best for real security

The two most frequently employed metrics in spoofing analysis are EER and t-DCF. EER is easy to understand; however, it makes the unrealistic assumption that false accepts and rejects have the same cost. AUC provides a broad picture but can minimize errors at critical phases, whereas FAR and FRR are threshold-dependent and subject to variation. Because it takes error costs into account, t-DCF is more practical for security; yet, it is complicated and primarily utilized in challenges. Although they are not as well studied, more recent metrics like minDCF and Precision–Recall may provide more insightful information.

6. Discussion

Safety authorities play a critical role in ensuring that Automatic Speaker Verification (ASV) systems are developed and deployed in a manner that prioritizes user privacy, security, and compliance with regulations. To mitigate new spoofing methods, law enforcement and regulatory agencies have initiated cross-sector task forces that work with ASV developers. This collaboration supports a proactive approach to identifying and countering spoofing threats [34], [35].
Overall, collaboration between data protection authorities, standardization bodies, consumer protection agencies, and law enforcement creates a comprehensive framework that enhances ASV systems' security and privacy.

6.1. Addressing Key Research Gaps in ASV Spoofing Detection

As the landscape of ASV systems evolves, it is essential to address the critical research gaps that hinder effective spoofing detection. This review paper also addresses significant unanswered questions and challenges in ASV spoofing detection, such as restricted generalizability, data variety, and adversarial attacks. Because of their heavy reliance on manually created features, classic models like GMMs frequently fall short against invisible spoofing types. In contrast, deep learning models (such as CNN, LSTM, and Transformers) exhibit greater robustness but require drastically larger datasets and processing capacity.
Our work focuses on advanced deep-learning techniques and addressing the ethical implications associated with spoof detection tools. Fairness and transparency are constrained by problems like dataset bias (such as poor speaker diversity across languages) and the "black-box" nature of deep models. The confidence of users can be increased by adopting explainable AI (XAI) strategies, such as post-hoc interpretation approaches or attention visualization.
Integrating these technical advancements with regulatory standards can ensure ASV technologies are not only effective but also aligned with privacy and security regulations [39].

6.2. Limitations and Implications for ASV Spoofing Detection

Difficulty in Handling All Attacks: Many models find it hard to detect different types of attacks, like fake or replayed voices, because they are trained on limited data. This makes them less useful in real-world situations.
High Resource Needs: Advanced models require a lot of computing power, making it hard to use them in devices with less processing ability or for quick detection. For instance, Wav2Vec and HuBERT-based systems achieve strong accuracy but are expensive in both training and inference compared to lightweight CNN or GMM models.
Newer Attack Methods: Attack methods are improving fast, and current systems often can't keep up, so we need solutions that can quickly adapt to new threats.

Evaluation Limitations: While common measurements like EER and t-DCF offer helpful baselines, they are insufficient to handle adjustments for factors like latency, energy efficiency, or demographic balance. A more detailed assessment of spoofing detection systems will be provided by merging ethical and computational parameters.

6.3. Strategy Followed for ASV Spoofing Detection

Our approach to addressing ASV spoofing challenges includes several strategies:

Limited Generalizability: New spoofing techniques and undiscovered data frequently cause models to falter. Through augmented training sets and novel model architectures, our approach achieves better adaptation to diverse spoofing types.

Data Diversity: It is necessary to collect large, diverse datasets that include a wide range of voices and spoofing techniques, but it can be costly and time-consuming. Future research could focus on multilingual and low-resource datasets to ensure global applicability.

Robustness against Adversarial Attacks: Adversarial attacks refer to techniques used by malicious actors to exploit vulnerabilities in ASV systems, often by creating deceptive inputs that mimic legitimate voices. By taking advantage of weaknesses, attackers might produce malevolent voices that bypass detection. We strengthen ASV models against potential spoofing exploits by employing adversarial training and regularization techniques [32].

Interpretability: Deep learning models can sometimes be "black boxes," which means it's challenging to figure out why a voice is identified as spoofing [40]. Techniques such as SHAP values, layer-wise relevance propagation, or saliency maps could improve interpretability.

Real-World Applications: Using models on devices with limited resources requires striking a balance between computing efficiency and accuracy [36]. For example, implementing ASV systems in mobile devices necessitates lightweight algorithms that maintain high accuracy without compromising performance, such as using model compression techniques. Edge-computing-oriented solutions and quantized neural networks are emerging directions here.

The following are some ways to overcome the issues faced in ASV spoofing

Emphasis on Spoofing Attacks: A major security problem raised in this study is the ASV systems' susceptibility to spoofing attacks. The lack of knowledge regarding the shortcomings of the ASV systems in use today and the requirement for reliable detection techniques is addressed by this.

Focused on Deep Learning Techniques: The study examines cutting-edge approaches, with a special emphasis on deep learning—a quickly developing subject that has promise for increasing the accuracy of spoofing detection. This fills the void of cutting-edge strategies to get around the drawbacks of conventional approaches.

7. Conclusion

Automatic Speaker Verification (ASV) is necessary for secure authentication, but it remains vulnerable to spoofing attacks like speech synthesis and voice conversion. In this review, we discussed ASV systems, commonly used datasets, evaluation metrics, and prevalent detection methods. We explain how deep learning enhances detection accuracy, strengthens its privacy, and adapts to new threats. Although methods like data augmentation and transfer learning help alleviate the problems of limited data, issues such as bias, explainability, and low-resource adaptation persist.

Furthermore, the computational efficiency of existing approaches is frequently problematic, which makes real-time deployment on devices with limited resources challenging.

Future research should also focus on incorporating privacy-preserving measures to guarantee ethical deployment and explainable AI approaches to boost transparency and confidence among users. This will make ASV systems more secure and trustworthy for use in real-world applications.

8. Statements and Declarations

Funding Information

No funds, grants, or other support were received.

Competing Interests

The authors declare that they have no conflict of interest.

Availability of Data and Materials

Not applicable

The authors have no relevant financial or non-financial interests to disclose.

Author Contribution

Conceptualization, methodology, formal analysis, writing—original draft preparation, writing—review and editing, Rekha Rani; supervision, Bal Kishan; All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We would like to sincerely thank all reviewers for their kind and constructive suggestions.

References

[1] A. Mittal and M. Dua, “Automatic speaker verification systems and spoof detection techniques: review and analysis,” *Int. J. Speech Technol.*, vol. 25, no. 1, pp. 105–134, Mar. 2022, https://doi.org/10.1007/s10772-021-09876-2.

[2] J. Boyd, M. Fahim, and O. Olukoya, “Voice spoofing detection for multiclass attack classification using deep learning,” *Mach. Learn. Appl.*, vol. 14, p. 100503, Dec. 2023, https://doi.org/10.1016/j.mlwa.2023.100503.

[3] M. Sajjad *et al.*, “CNN-based anti-spoofing two-tier multi-factor authentication system,” *Pattern Recognit. Lett.*, vol. 126, pp. 123–131, Sep. 2019, https://doi.org/10.1016/j.patrec.2018.02.015

[4] N. Shome, A. Sarkar, A. K. Ghosh, R. H. Laskar, and R. Kashyap, “Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges,” *Period. Polytech. Electr. Eng. Comput. Sci.*, vol. 67, no. 3, pp. 300–336, Jul. 2023, https://doi.org/10.3311/PPee.20971.

[5] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, “Voice spoofing detector: A unified anti-spoofing framework,” *Expert Syst. Appl.*, vol. 198, p. 116770, Jul. 2022, https://doi.org/10.1016/j.eswa.2022.116770

[6] E.-S. Atlam, M. Almaliki, G. Elmarhomy, A. M. Almars, A. M. A. Elsiddieg, and R. ElAgamy, “SLM-DFS: A systematic literature map of deepfake spread on social media,” *Alex. Eng. J.*, vol. 111, pp. 446–455, Jan. 2025, https://doi.org/10.1016/j.aej.2024.10.076.

[7] B. Chettri, “The Clever Hans Effect in Voice Spoofing Detection,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar: IEEE, Jan. 2023, pp. 577–584. https://doi.org/10.1109/SLT54892.2023.10022624.

[8] Y. Ren, H. Peng, L. Li, X. Xue, Y. Lan, and Y. Yang, “A voice spoofing detection framework for IoT systems with feature pyramid and online knowledge distillation,” *J. Syst. Archit.*, vol. 143, p. 102981, Oct. 2023, https://doi.org/10.1016/j.sysarc.2023.102981.

[9] J. Zhou, T. Hai, D. N. A. Jawawi, D. Wang, E. Ibeke, and C. Biamba, “Voice spoofing countermeasure for voice replay attacks using deep learning,” *J. Cloud Comput.*, vol. 11, no. 1, p. 51, Sep. 2022, https://doi.org/10.1186/s13677-022-00306-5.

[10] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, “Voice spoofing detection corpus for single and multi-order audio replays,” *Comput. Speech Lang.*, vol. 65, p. 101132, Jan. 2021, https://doi.org/10.1016/j.csl.2020.101132

[11] J. Guo, Y. Zhao, and H. Wang, “Generalized Spoof Detection and Incremental Algorithm Recognition for Voice Spoofing,” *Appl. Sci.*, vol. 13, no. 13, p. 7773, Jun. 2023, https://doi.org/10.3390/app13137773.

[12] R. Mohd Hanifa, K. Isa, and S. Mohamad, “A review on speaker recognition: Technology and challenges,” *Computer Electr. Eng.*, vol. 90, p. 107005, Mar. 2021, https://doi.org/10.1016/j.compeleceng.2021.107005

[13] H. Meng, W. Ou, J. Huang, H. Liang, W. Han, and Q. Zhang, “A robust unified spoofing audio detection scheme,” *Computer Electr. Eng.*, vol. 122, p. 109974, Mar. 2025, https://doi.org/10.1016/j.compeleceng.2024.109974.

[14] R. Rahmeni, A. B. Aicha, and Y. B. Ayed, “Speech spoofing countermeasures based on source voice analysis and machine learning techniques,” *Procedia Comput. Sci.*, vol. 159, pp. 668–675, 2019, https://doi.org/10.1016/j.procs.2019.09.222

[15] H. Liang, X. Lin, Q. Zhang, and X. Kang, “Recognition of spoofed voice using convolutional neural networks,” in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, QC: IEEE, Nov. 2017, pp. 293–297. https://doi.org/10.1109/GlobalSIP.2017.8308651.

[16] J. Monteiro, J. Alam, and T. H. Falk, “Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers,” *Comput. Speech Lang.*, vol. 63, p. 101096, Sep. 2020, https://doi.org/10.1016/j.csl.2020.101096.

[17] A. Chadha, A. Abdullah, L. Angeline, and S. Sivanesan, “A review on state-of-the-art Automatic Speaker verification system from spoofing and anti-spoofing perspective,” *Indian J. Sci. Technol.*, vol. 14, no. 40, pp. 3026–3050, Oct. 2021, https://doi.org/10.17485/IJST/v14i40.1279

[18] H. Tak, J. Patino, M. Todisco, et al., “End-to-end antispoofing with rawnet2,” in Proc. ICASSP, 2021, pp. 6369–6373. https://doi.org/10.1109/ICASSP39728.2021.9414234

[19] M. Barhoush, A. Hallawa, and A. Schmeink, “Speaker identification and localization using shuffled MFCC features and deep learning,” *Int. J. Speech Technol.*, vol. 26, no. 1, pp. 185–196, Mar. 2023, https://doi.org/10.1007/s10772-023-10023-2.

[20] M. A. Basit, C. Liu, and E. Zhao, “SDI: A tool for speech differentiation in user identification,” *Expert Syst. Appl.*, vol. 243, p. 122866, Jun. 2024, https://doi.org/10.1016/j.eswa.2023.122866

[21] School of Computer Science and Engineering, Taylors University, 47500, Selangor, Malaysia, A. Chadha, A. Abdullah, L. Angeline, and S. Sivanesan, “A review on state-of-the-art Automatic Speaker verification system from spoofing and anti-spoofing perspective,” *Indian J. Sci. Technol.*, vol. 14, no. 40, pp. 3026–3050, Oct. 2021, https://doi.org/10.17485/IJST/v14i40.1279

[22] S. Sinha, S. Dey, and G. Saha, “Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion,” *Comput. Speech Lang.*, vol. 86, p. 101599, Jun. 2024, https://doi.org/10.1016/j.csl.2023.101599

[23] Z. Almutairi and H. Elgibreen, “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions, “Algorithms 2022, 15, 155. https://doi.org/10.3390/a15050155.

[24] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, “Towards protecting cyber-physical and IoT systems from single- and multi-order voice spoofing attacks,” *Appl. Acoust.*, vol. 183, p. 108283, Dec. 2021, https://doi.org/10.1016/j.apacoust.2021.108283.

[25] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, “Advances in anti-spoofing: from the perspective of ASVspoof challenges,” *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. 1, 2020, https://doi.org/10.1017/ATSIP.2019.21

[26] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, “A study on data augmentation in voice anti-spoofing,” *Speech Commun.*, vol. 141, pp. 56–67, Jun. 2022, https://doi.org/10.1016/j.specom.2022.04.005.

[27] I. Himawan, F. Villavicencio, S. Sridharan, and C. Fookes, “Deep domain adaptation for anti-spoofing in speaker verification systems,” *Comput. Speech Lang.*, vol. 58, pp. 377–402, Nov. 2019, https://doi.org/10.1016/j.csl.2019.05.007

[28] A. T. Patil, H. A. Patil, and K. Khoria, “Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection,” *Comput. Speech Lang.*, vol. 72, p. 101301, Mar. 2022, https://doi.org/10.1016/j.csl.2021.101301.

[29] P. Gupta, H. Patil, and R. Guido, “Vulnerability issues in Automatic Speaker Verification (ASV) systems,” EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:10, https://doi.org/10.1186/s13636-024-00328-8.

[30] P. Abdzadeh and H. Veisi, “A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection,” *J. AI Data Min.*, vol. 11, no. 1, Jan. 2023.

[31] J. Xue, H. Zhou, H. Song, B. Wu, and L. Shi, “Cross-modal information fusion for voice spoofing detection,” *Speech Commun.*, vol. 147, pp. 41–50, Feb. 2023, https://doi.org/10.1016/j.specom.2023.01.001.

[32] X. Wang and J. Yamagishi, “A comparative study on recent neural spoofing countermeasures for synthetic speech detection*,” in Proc. Interspeech*, 2021, pp. 4259– 4263. https://doi.org/10.21437/Interspeech.2021-702

[33] L. Nguyen, M. Bui et al., “On the Defense of Spoofing Countermeasures Against Adversarial Attacks,” *IEEE Access*, date of publication 31 August 2023, Digital Object Identifier https://doi.org/10.1016/j.jksuci.2022.02.024.

[34] H. Dawood, S. Saleem, F. Hassan, and A. Javed, “A robust voice spoofing detection system using novel CLS-LBP features and LSTM,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7300–7312, Oct. 2022, https://doi.org/10.1016/j.patrec.2011.06.011.

[35] Y. W. Wong *et al.*, “A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities,” *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1503–1510, Oct. 2011, https://doi.org/10.1016/j.patrec.2011.06.011.

[36] R. Rahmeni, A. B. Aicha, and Y. B. Ayed, “Acoustic features exploration and examination for voice spoofing countermeasures with boosting machine learning techniques,” *Procedia Comput. Sci.*, vol. 176, pp. 1073–1082, 2020, https://doi.org/10.1016/j.procs.2020.09.103.

[37] Y. Huang, Q. Shen, and J. Ma, “AFP-Conformer: Asymptotic feature pyramid conformer for spoofing speech detection,” *Speech Commun.*, vol. 166, p. 103149, Jan. 2025, https://doi.org/10.1016/j.specom.2024.103149.

[38] P. Ziabari and H. Veisi, “A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection, “*Journal of Artificial Intelligence and Data Mining (JAIDM),* Vol. 11, No. 1, 2023, 119-129.

[39] M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, “Speaker identification security improvement by means of speech watermarking,” *Pattern Recognit.*, vol. 40, no. 11, pp. 3027–3034, Nov. 2007, https://doi.org/10.1016/j.patcog.2007.02.016.

[40] B. Wickramasinghe, E. Ambikairajah, V. Sethu, J. Epps, H. Li, and T. Dang, “DNN controlled adaptive front-end for replay attack detection systems,” *Speech Commun.*, vol. 154, p. 102973, Oct. 2023, https://doi.org/10.1016/j.specom.2023.102973.