Unmasking the Audio Illusion: A Survey on Spoofing and Deepfake Detection

Aarthi S and Akshay Agarwal
Trustworthy BiometraVision Lab, IISER Bhopal, India
{saarthi24,akagarwal}@iiserb.ac.in

Abstract

With breakthroughs in deep learning algorithms, the practice of manipulating audio to produce believable fakes is expanding rapidly. This survey paper provides a comprehensive overview of the current state of deepfake audio research, encompassing generation methods, online platforms to generate fake audio, the latest detection techniques, human perception of fake audio, and the underlying security concerns. We examine different methods for speech synthesis, audio splicing, and voice cloning, pointing out their advantages and disadvantages. Furthermore, we investigate various detection algorithms, encompassing supervised, unsupervised, and hybrid techniques, and assess their effectiveness in detecting audio manipulation. We review deepfake audio's impacts, including possible adverse effects on reputation, fraud, and misinformation. We present a concise analysis of AI versus human detection of deepfake audio, drawing insights from existing literature and validating them through our experiments. Finally, we highlight future research directions and recommendations for mitigating the societal risks associated with this powerful technology.

1. Introduction

Audio is extensively utilized for Automatic Speaker Verification Systems (ASVs) and Personal Voice Assistants (PVAs). For instance, HSBC uses ASV, implementing "Voice ID" in 2019, allowing customers to access certain banking services such as account balances and transfers through phone banking simply by using their voice [10]. PVAs such as Google Home, Alexa, and Cortana have become helpful companions in our homes, streamlining daily tasks with simple voice commands. The global voice biometrics market was valued at 1467.4 US million dollars in 2022 and is projected to reach 4985.8 US million dollars by 2030, at a CAGR of 19.1% during the forecast period [49]. The above statistics reflect the growth and importance of the voice/speech/audio modality and its presence in our day-to-day lives. However, studies show that voice-based systems

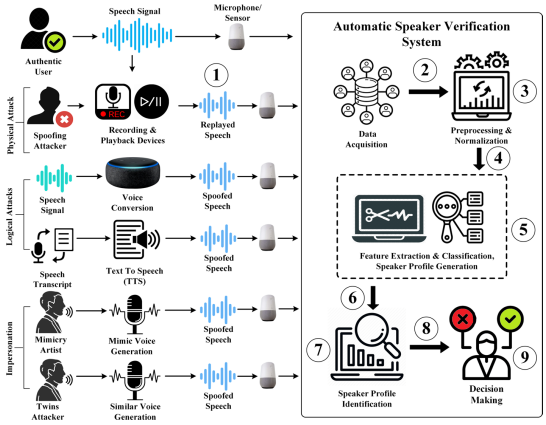

Figure 1. Showcasing possible attacks on automatic speaker verification systems [23].

are susceptible to various attacks, including voice cloning and deepfakes, as shown in Figure 1. These attacks span the entire pipeline: from injecting fake audio signals at the microphone, tampering during signal preprocessing and feature extraction, to manipulating the matcher and decision modules, ultimately enabling unauthorized access even with minimal authentic input.

Audio deepfakes involve modifying the original voice recordings to impersonate someone or using someone's voice to create new utterances. Significant progress has been made in deepfake generation techniques, including text-to-speech, voice encoders, voice conversion, and voice cloning. Due to these advancements, fooling machines and humans is an easy task; several real-world cases are proof of that. For example, a 73-year-old man from Kerala, India, received a call in July for 9, 2023, from an anonymous caller impersonating his former colleague and asking for money. The victim transferred a sum of $\mathbf{40,000}$ rupees before realizing he had been tricked [48]. Large Language Models (LLMs) are also vulnerable to adversarial audio attacks [63]. Further, the strength and advancement of such artificial intelligence (AI) deepfake can be seen from the *political incidents, which are seen as a potential attempt to interfere with the US national election*[1]. These incidents

[1]https://www.resemble.ai/deepfake-incident-joe-biden-election-

highlight the growing concern about fake technology being used maliciously. Interestingly, the technology has reached the level that performing a voice deepfake or cloning is now available at a click of a button that does not involve any technical knowledge for a novice user[2].

The escalating threat of fake voice attacks has triggered a surge in research. In 2024 alone, a staggering 4248 papers were published on audio deepfakes, compared to only one paper published in 2017. This survey primarily focuses on audio attacks, audio deepfake generation, and detection capabilities of various deep learning models, ChatGPT4.0, and humans, which are ignored in the majority of previous surveys [23, 27], in addition to assessing their current state and how to lessen their negative consequences. In the end, we have also performed extensive experimental analysis demonstrating the strength of human, AI, and large foundation models in detecting audio deepfakes. We assert that the knowledge of the literature and the experimental validations make this research a unique contribution compared to generic surveys.

2. Audio Synthesis Attacks: Taxonomy and Techniques

This section provides a taxonomy of the attacks and describes popular techniques for fake audio generation.

2.1. Attacks Taxonomy

In the context of audio, attacks encompass misinformation, privacy breaches, and security threats. Manipulated audio can spread false information, erode trust, and influence public opinion. Privacy concerns arise from voice cloning and unauthorized eavesdropping, while security implications include identity theft, fraud, and disruption of communication channels. Addressing these challenges requires a comprehensive approach involving technological advancements, legal frameworks, and heightened public awareness.

In the pursuit of practical audio attacks, four key ideal properties emerge. (i) Firstly, attacks should be **"Over-the-Air"**, involving the transmission of adversarial examples through loudspeakers, presenting a real-world challenge due to distortions introduced by device characteristics, channel effects, and ambient noise. (ii) Secondly, attacks should be adaptable to **"Black Box Systems"**, considering the limited knowledge of commercial voice assistants. Query-based optimization methods, transferability between models, and substitute models contribute to attacking black-box systems effectively. (iii) Thirdly, attacks should maintain **"Imperceptible Adversarial Perturbations"** by optimizing amplitude, frequency, temporal alignment, and noise features. These optimizations aim to reduce noticeability and improve the stealthiness of adversarial examples. (iv) Lastly, attacks should be capable of **"Real-Time Execution"**, requiring on-site generation or adjustment. The presence of these four properties signifies a greater practical threat in real-world audio security scenarios.

Further, fake audio attacks can be classified into three categories: logical attacks, which include speech synthesis and voice conversion; physical attacks, such as replay and impersonation; and adversarial attacks, which involve manipulating audio signals to deceive or mislead audio processing systems by introducing imperceptible perturbations into the audio data, leading to misinterpretations by the targeted system [45].

2.2. Fake Audio Generation Methods

Creating compelling deepfake audio relies on various techniques utilizing deep learning models. Here are some prominent approaches:

Voice conversion: Voice cloning encompasses various methods such as autoencoders, generative adversarial networks (GANs), transfer learning, and Mel Spectrogram Inversion. Recent developments in voice cloning have focused on improving the expressiveness and naturalness of synthesized speech [2], mainly when limited training data is available. Voice conversion can be further improved by combining a speaker's voice with the mimicking of prosody, achieving high quality and similarity to the original voice and prosody [34]. Nowadays, several online services offer AI-enabled voice cloning, as mentioned in Table 1.

Text-to-speech (TTS) Synthesis: TTS has evolved significantly over the years, focusing on creating natural-sounding and expressive speech. This noteworthy development opens up new possibilities for voice assistants and audiobooks, making immersive audio experiences easier. Several end-to-end models, such as FastDiff-TTS and Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS), have been presented to produce high-quality speech.

Beyond replication, the scope of TTS extends to the domain of expressiveness. GAN-based models such as GAN-TTS and MelGAN [26] provide unprecedented control over pitch, intonation, and emotional subtleties. Still, the complexities of human emotion, with all of its subtle pauses and turns, present a constant difficulty. Researchers are currently delving into the intricacies of prosody and style transfer to smoothly incorporate language and emotion into synthetic speech and provide an authentic impact that appeals to listeners. Style tokens and reference encoders have been used to improve expressive TTS, while prosody prediction techniques, including duration and pitch modeling, further enhance the naturalness of speech. The concept of

interference/

[2]https://speechify.com/voice-cloning/

Online-toolbox	Description
Amphion [69]	Supports generation tasks such as Text to Speech, Text to Audio, and Singing Voice Conversion.
VALL-E [53]	Neural codec language-model for generating speech.
SpeechSplit 2.0 [1]	Unsupervised voice conversion by speech disentanglement using multiple autoencoders.
Descript Overdub	Descript's TTS generator transforms any text into natural-sounding speech in minutes.
Resemble AI	Popular AI voice generator and robust deepfake audio detection tool
iSpeech	Offers both TTS and voice recognition capabilities.
Tacotron2 and WaveNet	Speech synthesis by generating waveforms that closely match natural human speech.

Table 1. Popular online toolbox for audio deepfake generation.

personalized voice generation in TTS using speaker embeddings helps create voices that match specific visual stimuli [52]. Additionally, an emotion-vector-based synthesis method enables computationally efficient manipulation of primary emotions, polarity, and intensity levels while ensuring high naturalness and transferability across unseen speakers [19]. Also, pre-trained text embeddings, such as BERT, have been proposed to improve the naturalness of synthesized speech [17]. Recent advancements continue to redefine the boundaries of TTS. F5-TTS [68] introduces a fully non-autoregressive framework using flow matching with a Diffusion Transformer (DiT), achieving high naturalness, zero-shot expressiveness, and code-switching efficiency without requiring duration modeling or phoneme alignment. Llasa [66] brings the scalability of large language models (LLMs), particularly LLaMA, to speech synthesis, demonstrating how scaling train-time and inference-time compute enhances prosody and naturalness through a simplified architecture using a single VQ codec and Transformer. Similarly, MaskGCT [56] adopts a masked prediction approach across two stages—semantic and acoustic token generation—allowing fully non-autoregressive zero-shot TTS without alignment or duration supervision, thus simplifying the synthesis pipeline while improving flexibility and performance.

Replay Attack: A replay attack involves recording a target speaker's voice and replaying it to impersonate them, often to bypass voice authentication systems. Techniques associated with replay attacks include far-field detection, where the attacker plays back a recording through a phone or speaker, and audio splicing (cut-and-paste), where a fake utterance is constructed by splicing together short recordings. A key example of replay attacks is the ASVspoof challenge, held annually since 2015, focusing on automatic speaker verification spoofing and countermeasures. In the latest 2024 challenge, the authors have presented a combined logical access and deepfake task comprising challenging crowd-sourced data while also incorporating adversarial attacks [55].

Apart from the attacks mentioned above, a range of studies have explored the vulnerability of audio systems to adversarial attacks. The overall framework of the generation of deepfake audio involves a meticulously crafted pipeline, which can be summarised as below:

1. **Data Acquisition and Preprocessing:** (i) Gather audio recordings of the individual whose voice will be mimicked. (ii) Remove noise, unwanted segments, and silence from the recordings. Segment the remaining audio into smaller units, such as phonemes or sentences, for easier processing. (iii) Extract relevant acoustic features from the audio data. This can include Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and pitch information.
2. **Model Training:** (i) For text-to-speech (TTS) systems, the model is trained with paired text and audio data. The model learns to generate speech audio from the input text by predicting the corresponding acoustic features of the target speaker's voice. (ii) For voice conversion, the model is trained to map the acoustic features of the source speaker's voice to those of the target speaker, effectively transforming the voice while maintaining the linguistic content.
3. **Deepfake Generation:**
   - Text Input: Provide the desired text as input to the trained model for text-to-speech deepfakes.
   - Speaker Embedding: Generate a compressed representation of the target speaker's voice, capturing their unique vocal characteristics. This often involves passing the preprocessed audio data through an encoder network.
   - Mel Spectrogram Synthesis: Combine the speaker embedding with text input to generate a Mel Spectrogram, representing the harmonic structure and rhythm of the desired speech using models such as Tacotron 2.
   - Waveform Generation: Convert the Mel Spectrogram into the actual audio waveform using a vocoder model such as WaveNet. This final stage transforms the abstract representation into a realistic-sounding voice.
4. **Post-processing and Refinement:**

- Smoothing and Polishing: Apply denoising and smoothing techniques to remove artifacts and enhance the naturalness of the generated audio.
- Background Noise Addition: Add background noise if required to match the intended context of the deepfake audio.

5. **Output and Deployment:** The final deepfake audio is ready for use. It can be integrated into videos, used for voice manipulation, or incorporated into various creative applications.

3. Fake Voice Detection

The sense of relief comes from the fact that while tremendous effort has been put into generating fake audio, a separate school of thought is actively working on detecting it. However, before developing a unified and effective fake audio detector, it is necessary to have a dataset that can reflect a variety of phony audio. A comprehensive list of existing fake voice detection datasets and some multimodal datasets, such as HAV-DF and DefakeAVMiT, is reported in Table 2. The Table highlights the lack of freely available datasets encompassing diverse languages, ethnicities, and real-world scenarios.

Utilizing the strength of existing datasets, several research efforts have started to develop an effective and accurate fake voice detection architecture. For example, Utilization of a pre-trained wav2vec2 feature extractor and a downstream classifier [35] as well as employing a 34-layer ResNet with multi-head attention pooling and neural stitching [61] for effective detection of audio deepfakes. Further, we comprehensively survey the existing voice/audio deepfake detection algorithms by categorizing them into multiple classes based on the ML concepts they used, such as traditional feature-based to advanced deep network architecture.

3.1. Handcrafted Feature-Based Algorithms

Perceptual features such as Perceptual Linear Prediction (PLP) and Constant-Q Cepstral Coefficients (CQCC) are highly significant, showing promising results in the Audio Deep Synthesis Detection (ADD) Challenge [28]. Approaches using Fundamental Frequency Variation (FFV) and spectral features [6, 15, 29] also play a key role. In place of classifiers utilizing single discriminating information, several research efforts have also been proposed utilizing multimodal information. The frequency characteristics of the audio signal can be analyzed to identify unnatural patterns or inconsistencies that may indicate spoofing. These methods are built end-to-end, proposing a discriminative frequency information SincNet for speech antispoofing [32]. Further, a multimodal approach of integrating audiovisual features for deepfake detection enhances intra- and cross-domain testing performance [41, 64].

3.2. End-to-end Deep Learning-Based Models

While the handcrafted feature-based techniques are practical and cost-efficient, their resilience to unseen voice deep-fakes is a critical drawback. Therefore, by looking at the generalization capacity of deep networks, a shift in voice deepfake detection algorithms has been noticed that primarily utilize deep learning architectures. For example, Convolutional Neural Networks (CNNs) can be used for audio deepfake detection by focusing on audio recapture detection [25]. The study demonstrates the effectiveness of CNNs in capturing subtle patterns and inconsistencies in deepfakes. A fully automated system combines a modified version of Differentiable Architecture Search (DARTS) and a pre-trained model to achieve superior performance [54]. A vision transformer network can also classify spectrogram images and detect deepfake audio [51]. These studies collectively demonstrate the effectiveness of deep learning in detecting fake audio. The above review shows that combining techniques of different domains, exploring new features and representations, and addressing challenges such as real-time detection are key areas for future advancements.

In addition to these efforts, recent deep learning-based models, such as AVA-CL [70], demonstrate excellent cross-manipulation detection performance, achieving an AUC of 97.89%. However, AVA-CL faces audio inconsistencies when fake videos include facial flickering. ResNet-101SV also highlights the importance of fine-tuning diverse datasets, achieving impressive EER scores on ASVspoof and wild datasets. However, the absence of advanced augmentation techniques limits its robustness [4]. Models such as ASDG [57] and SLIM have made significant achievements in domain generalization but are still limited by dataset constraints and potential misclassification risks [72]. The major limitation observed is the generalizability of these algorithms, as illustrated in Figure 2. Figure 2 indicates that the performance of models trained for a particular dataset fails miserably for out-of-distribution datasets. For example, popular models such as Transformer, WavLM, and Whisper show an EER of 7.50%, 7.24%, and 5.59%, respectively, on the ASVSpoof dataset. Still, the EER rises significantly to 43.78%, 30.50%, and 42.73%, respectively, on the in-the-wild dataset. This performance drop is even more drastic in traditional detectors such as LCNN, which exhibit huge increases in EER when evaluated on out-of-distribution sets such as LibriSeVoc and In-the-Wild, compared to their in-domain performance. In contrast, foundation models such as Hubert and Whisper demonstrate greater resilience. This robustness stems from their large-scale pretraining. Larger model capacities and rich feature abstractions allow these architectures to capture nuanced ar-

Dataset	Language	Samples (S)/ Hours (H)	Key Features	Limitations	Availability
InDeepFake [5]	7 languages	5069	Multimodal Indian deepfake video dataset covering multiple age and gender groups.	Lacks robustness under adversarial attack scenarios.	Public
JMAD [36]	17 languages	412,021	Covers TTS, vocoder, voice-conversion, and adversarial attacks.	Exhibit wide MOS variability, introducing inconsistencies that can affect training and benchmarking.	Private
ASVspoof5 [55]	English	1,500,713	Encompass 32 spoofing algorithms—including TTS, VC, and novel adversarial attacks.	Focuses solely on the English subset of MLS.	Public
EMILIA [18]	6 languages	101000 H	Captures diverse acoustic and semantic speaking styles of real, spontaneous human speech.	Lack precise alignment between text and audio or fine-grained emotion annotations.	Public
MLAAD [39]	38 languages	378 H	Synthetic audio generated using 82 TTS models comprising 33 different architectures.	Lacks coverage of voice conversion attacks and real-world noise conditions.	Public
MSTF [3]	Caucasian, Black, South Asian, and East Asian	143754	Multi-scenario talking face dataset, featuring 22 audio and video forgery techniques.	Lacks robustness to compression artifacts of social media platforms.	Private
HAV-DF [20]	Hindi	500	Hindi multimodal dataset encompassing face-swapping, lip-syncing, and voice-cloning manipulations.	Small dataset with limited manipulation techniques.	Private
DeepFakeVox-HQ [71]	English	1300000	Large voice dataset covering various AI voice synthesis models and social media platforms.	Limited to particular types of corruption and adversarial attacks.	Private
FakeSound [59]	English	3798	Identifying manipulated audio and locating deepfake segments using grounding, masking, and inpainting.	Struggles with domain adaptation, limiting its generalizability to unseen domains.	Public
LlamaPartialSpoof [33]	English	130 H	Full and partial fake speech, using LLM and voice cloning to evaluate the robustness of countermeasures.	Does not address real-world conditions such as noisy or reverberant environments.	Public
VoiceWukong [62]	English, Chinese	413400	Generated using 19 commercial and 15 open-source tools encompassing 38 variants of six manipulation types.	Lacks representation of other widely spoken languages.	Public
Codecfake [58]	English, Mandarin	1000000	Designed to detect novel Audio Language Model (ALM)-based deepfake audio.	The Dataset is limited to speech alone.	Public
Cross-Domain [31]	English	300 H	To detect advanced zero-shot text-to-speech models that can clone voices from a single utterance.	Neural codec compressors pose accuracy challenges.	Public
TIMIT-TTS [46]	English	80000	Created using TTS and Dynamic Time Warping (DTW) for multimodal synthetic media detection.	Detection accuracy degrades due to post-processing operations, and for single-speaker systems.	Private
DefakeAVMiT [64]	English	7020	Multimodal dataset where both audio and visual modalities can be independently forged using eight different techniques.	Lip-syncing techniques reduce explicit forgery traces.	Private

Table 2. Latest existing datasets for deepfake audio detection.

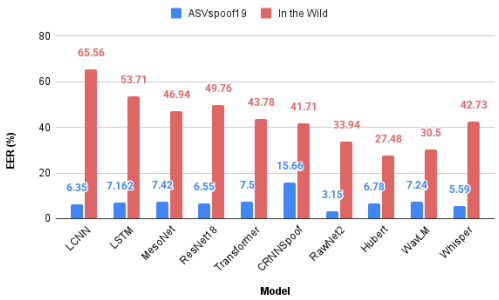

Figure 2. Showcasing how well the existing algorithms trained on ASVSpoof19 perform when applied to an in-the-wild dataset [38, 65]. The high error rate reflects the current limitation. The results of the best-performing network-feature combination are taken from [65].

tifacts that handcrafted features and shallow models often miss [30]. Not only generalization but also the adversarial robustness of the existing algorithms is a serious issue that has not been adequately addressed for audio compared to images [9, 22, 67]. These further emphasize the importance of improving model robustness, data diversity, and scalability for efficient detection of audio deepfakes.

4. Human vs. ChatGPT4.0 vs Audio Spectrogram Transformer (AST) in Audio Deepfake Detection

To complement the survey of existing deepfake detection techniques, this section presents a concise experimental analysis comparing the detection capabilities of humans, the Audio Spectrogram Transformer (AST), and ChatGPT4.0. This empirical exploration is further supported by relevant literature, offering practical insights into how well current systems and humans can detect synthetic audio. We curate a balanced subset of 200 audio samples (100 real and 100 fake) from the FakeAVCeleb multimodal dataset, ensuring balanced representation across all ethnicities and genders. By integrating human perceptual judgments, advanced transformer-based models, and language-model-based reasoning, this analysis bridges theory with real-world detection performance, revealing nuanced differences across evaluators.

4.1. Human Examiners

We survey multiple participants encompassing diverse educational backgrounds, domain expertise, and gender through Google Forms. Initially, participants are randomly asked to identify the audio provided as "real" or "fake", and this performance is recorded as "Without Knowledge Score". Then, they are asked to listen to real and fake audio samples (precisely five samples of each class randomly selected with enough class representation) provided with appropriate labels as a training phase. After training, they are

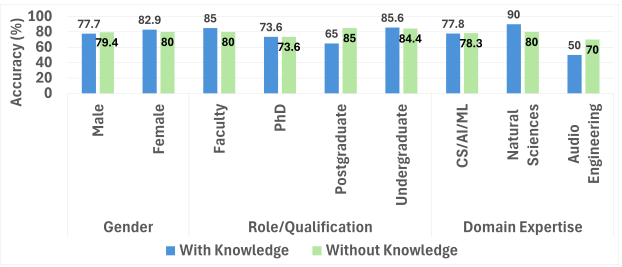

Figure 3. Performance analysis based on a human survey of audio deepfake detection.

again asked to detect the same audio as "real" or "fake", which is reported as "With Knowledge Score". However, it is observed that this training process doesn't necessarily improve performance. Lack of emotion, unnatural breathing, and poor audio quality are commonly reported by participants as reasons to identify audio as fake. Generally, humans tend to show above-average performance in the given dataset, as illustrated in Figure 3, indicating the need for a better deepfake audio dataset. The figure also illustrates that gender and domain expertise play little role in audio deepfake detection.

Research on human perception of audio deepfakes has revealed various factors affecting detection accuracy and performance. Evidence [40] indicates that AI-based models can outperform human participants in detecting audio deepfakes, particularly under unrealistic conditions where models overfit the training dataset. However, in more realistic scenarios (where shortcuts, such as the length of audio silence, are removed), humans and AI models perform similarly, sharing common strengths and weaknesses. This highlights that while AI can exploit dataset-specific artifacts, its real-world performance converges with human capabilities when tested in more dynamic environments.

There are findings [16] that second this by showing that human observers consistently perform worse than SOTA AI models when identifying audiovisual deepfakes. Interestingly, the study also finds that after placing a deepfake, participants often struggle to pinpoint the exact manipulation (audio or video or both), particularly mistaking changes in the audio for changes in the visuals. Furthermore, being alerted about the presence of deepfakes or increasing familiarity with them had little effect on detection accuracy, suggesting that training alone has a limited impact on improving human detection skills. In another study [14], researchers find that humans rely more on audiovisual cues than the content itself when detecting deepfake political speeches. Specifically, TTS-based deepfakes are much more challenging for humans to detect than voice conversion or waveform concatenation systems. Similarly, a cross-country analysis [11] reveals that across all countries (USA, Germany, and China) and media types (image, audio, and text), participants struggle to distinguish between real and

AI-generated media. Participants often rate artificially generated media as human-made, performing worse than random guessing in some cases. The study emphasizes that machine-generated media are becoming virtually indistinguishable from real media and that human participants tend to rely on irrelevant or misleading cues, further complicating the detection process. In addition, proficiency in native language, age, and exposure are key factors that affect human performance in detecting audio deepfakes, while technical experience shows little influence [14, 16, 40]. All of these factors pave the way for future research.

4.2. AI-based Experimental Analysis

We extend our study on audio deepfake detection by performing experiments using AST and ChatGPT's audio analyzer on the same set of 200 audios used in our human survey.

- **ChatGPT Audio Analyzer:** The 200 audios are tested with ChatGPT's audio analyzer using the prompt, 'Analyze the given audio and comment if it is real or fake based on your analysis.' The accuracy is found to be 49%, with ChatGPT failing to recognize most of the fake audio. ChatGPT detects fake audio based on attributes such as frequency and amplitude range. An intriguing study on deepfake detection using ChatGPT [47] concludes that ChatGPT can detect multimodal deepfakes comparable to humans given suitable prompts and context. However, multimodal ChatGPT's performance is quite poor compared to other AI models trained exclusively for deepfakes [47]. Without specialized multimodal models, its performance lags behind state-of-the-art deep learning models optimized for audio and audiovisual forgery detection.
- **AST:** The pre-trained AST is fine-tuned with the same 200 audios used for detection by humans and ChatGPT using a 50:50 train-test split as well as 3-fold cross-validation, and every time, the accuracy is found to be 100%. AST outperforms ChatGPT and humans in audio deepfake detection, as illustrated in Figure 4. While the above analysis showcases promising results through AST and humans, their collaboration can be explored in the future, where a website plugin can be developed, using human understanding and AI knowledge to counter audio deepfakes effectively.

5. Prevention and Future Directions

Detecting audio deepfakes can significantly contribute to rebuilding trust in society in digital data. In this section, we provide possible research directions that can be used to prevent audio deepfakes. Addressing these limitations in future defenses could pave the way for a robust and unified system against audio deepfakes, offering significant real-world applications.

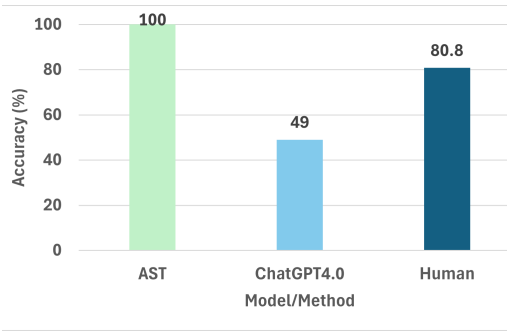

Figure 4. Comparison of human, ChatGPT, and AST performance in audio deepfake detection. Note: the high performance of AST must not be treated as a complete solution to the problem due to the limited sample size.

5.1. Prevention

- **Liveness Detection:** In a replay attack, voices are generated by a live person but replayed on some electronic media. One preventive measure to avoid an attack is identifying whether a living human generates the voice sample. For example, pop noise caused by human breath can be used as a liveness indicator [50]. Another way to prevent replay of the recorded media is to pose real-time questions. If the media is not a living person, it would be challenging for an attacker to solve such real-time questions.
- **Behavioral Analysis:** The behavioral patterns of users can be monitored during audio interactions that could suggest fraudulent activity. Audio and visual modalities can be simultaneously analyzed along with perceived emotions, achieving high accuracy in deepfake detection [37].
- **Use of Multiple Cues:** Multiple authentication methods or cues, such as voice and facial recognition, can be combined to create a robust system. Combining different biometric modalities, such as face and speaker identification, can significantly improve fake detection accuracy [8].
- **Blockchain Technology:** Blockchain can secure both deep learning models (used for fake audio generation) and audio content against tampering and misuse. Approaches such as Deepring store model parameters and metadata immutably, enabling integrity verification and controlled access [12, 13]. This prevents unauthorized audio synthesis and ensures traceability of genuine content. Deploying detection in high-stakes workflows requires balancing on-chain transparency with off-chain scalability and privacy, especially under GDPR and deepfake-labeling laws [44].

- **Regulatory Measures:** Advocate for and comply with regulatory measures that address deepfake creation and distribution, promoting legal consequences for malicious use [7].
- **Others:** There are several ways to prevent the spread of fake content. Educating users, particularly those unaware of the technology's ability to generate human-like content, is crucial [43]. Verifying unreliable content through digital signatures or watermarking can help trace creators. Additionally, developing diverse datasets across languages and demographics can aid in building automated detection algorithms. These solutions, if refined, could be made accessible as browser extensions or mobile apps.

5.2. Open Challenges: Future Directions

This section focuses on the existing challenges and research gaps in audio deepfake detection.

- **Data Availability and Diversity:** Limited access to diverse datasets of real and fake audio, encompassing various speaker demographics, languages, and attack techniques, hinders the development of robust and generalizable detection algorithms. As listed in Table 2, the existing fake audio datasets target limited languages or are highly biased toward English and Caucasian ethnicity.
- **Generalization Robustness:** Audio deepfakes use techniques such as voice cloning, voice conversion, or synthesis from scratch. Limited evaluation of detection algorithms in real-world scenarios, including unseen attacks, creates false security. Developing models that generalize across attack types and ensuring robust assessment are crucial for effective defense [21].
- **Privacy Concerns:** Deploying large-scale audio monitoring systems for spoofing detection raises privacy concerns regarding data collection and usage. Balancing security with individual privacy is crucial.
- **Performance on Limited Data and Zero-Shot Scenarios:** Existing challenges persist in achieving robust deepfake audio detection, particularly when confronted with limited training data and encountering entirely new voices (zero-shot scenarios) [24]. While many current algorithms demonstrate high accuracy in controlled datasets, their effectiveness falters under a broad spectrum of voices with minimal labeled data, as clearly evident from Figure 2.
- **Interpretability and Explainability:** Current deep learning models for deepfake detection often lack interpretability, making it difficult to understand why specific audios are classified as real or fake. This hinders building trust and confidence in the system.
- **Computational Cost:** Implementing complex deep models for real-time detection on resource-constrained devices such as smartphones remains computationally expensive. Noteworthy is that these smartphones are one of the primary victims of fake content.
- **Language Barrier:** Current speaker verification systems, used for authentication in various languages, are vulnerable to deepfakes. Each language requires a separate model, which is impractical and expensive due to the time and computational resources needed for training and deployment. Therefore, developing language-independent models would be a breakthrough, significantly improving the efficiency and effectiveness of deepfake detection systems.
- **Multimodal Detection:** Deepfake content often involves multimodal manipulation; hence, integrating multimodal detection techniques can better help detect fake content [42].
- **Unintended Bias:** Deepfake detection systems may unintentionally exhibit biases, leading to false positives or false negatives. Ensuring fairness and minimizing unintended bias in detection models, especially across diverse demographic groups, is an important ethical consideration [38, 60].

6. Conclusion

The rapid evolution of deepfake audio generation methods poses a significant challenge to the integrity of audio content, with potential consequences for various applications, from voice cloning to voice authentication systems. Researchers have made commendable progress in developing detection methodologies in response to the escalating threat of audio deepfakes. However, despite these advancements, several challenges persist. We discuss the unmet research needs and unresolved issues that require careful consideration and ongoing investigation. Building reliable audio spoofing detection systems requires resolving or offering answers to these open research issues. Collaboration between academia, industry, and regulatory bodies, along with continuous education and awareness initiatives, will be essential to stay ahead of the evolving landscape of audio deepfakes and safeguard the integrity of audio content in various domains. In contrast to the existing survey papers, this paper has provided an in-depth exploration of the diverse landscape surrounding deepfake audio, covering generation, detection, human perception, prevention strategies, and the effectiveness of various stakeholders such as humans, AI, and foundational models.

Acknowledgement

Akshay Agarwal is partially supported through the PM-ECRG grant of ANRF. The INSPIRE fellowship of DST, India, partially supports Aarthi S.

References

[1] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson. Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6332–6336. IEEE, 2022. 3

[2] B. Chen, C. Du, and K. Yu. Neural fusion for voice cloning. *IEEE-ACM TASLP*, 30:1993–2001, 2022. 2

[3] X. Chen, Q. Yin, J. Liu, W. Lu, X. Luo, and J. Zhou. Glcf: A global-local multimodal coherence analysis framework for talking face generation detection. *arXiv preprint arXiv:2412.13656*, 2024. 5

[4] A.-T. Dao, M. Rouvier, and D. Matrouf. Asvspoof 5 challenge: advanced resnet architectures for robust voice spoofing detection. In *Proc. ASVspoof 2024*, pages 163–169, 2024. 4

[5] A. K. Das, A. Bose, P. Manohar, A. Dutta, R. Naskar, and R. S. Chakraborty. Indeepfake: A novel multimodal multilingual indian deepfake video dataset. *Pattern Recognition Letters*, 2025. 5

[6] J. Deng, T. Mao, D. Yan, L. Dong, and M. Dong. Detection of synthetic speech based on spectrum defects. In *DDAM*, pages 3–8, 2022. 4

[7] Express News Service. New regulations to tackle deepfakes soon: It minister vaishnaw. https://indianexpress.com/article/india/new-regulation-deepfakes-soon-vaishnaw-social-media-platforms-9039093/, Nov. 2023. Accessed: 2025-05-05. 8

[8] M.-I. Faraj and J. Bigun. Synergy of lip-motion and acoustic features in biometric speech and speaker recognition. *IEEE ToC*, 56(9):1169–1175, 2007. 7

[9] M. U. Farooq, A. Khan, K. Uddin, and K. M. Malik. Transferable adversarial attacks on audio deepfake detection. *arXiv preprint arXiv:2501.11902*, 2025. 6

[10] E. Flitter and S. Cowley. Voice deepfakes are coming for your bank balance. https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html, 2023. 1

[11] J. Frank, F. Herbert, J. Ricker, L. Schönherr, T. Eisenhofer, A. Fischer, M. Dürmuth, and T. Holz. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 55–73. IEEE, 2024. 6

[12] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Deepring: Protecting deep neural network with blockchain. In *IEEE/CVF CVPRW*, 2019. 7

[13] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Securing cnn model and biometric template using blockchain. In *IEEE BTAS*, pages 1–7, 2019. 7

[14] M. Groh, A. Sankaranarayanan, N. Singh, D. Y. Kim, A. Lippman, and R. Picard. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature communications*, 15(1):7629, 2024. 6, 7

[15] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10:134018–134028, 2022. 4

[16] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang. Unmasking illusions: Understanding human perception of audiovisual deepfakes. *Authorea Preprints*, 2024. 6, 7

[17] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu. Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis. In *Interspeech 2019*, pages 4430–4434, 2019. 3

[18] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024. 5

[19] P. Kalyan, P. Rao, P. Jyothi, and P. Bhattacharyya. Emotion arithmetic: Emotional speech synthesis via weight space interpolation. In *Proc. Interspeech 2024*, pages 1805–1809, 2024. 3

[20] S. Kaur, M. Buhari, N. Khandelwal, P. Tyagi, and K. Sharma. Hindi audio-video-deepfake (hav-df): A hindi language-based audio-video deepfake dataset. *arXiv preprint arXiv:2411.15457*, 2024. 5

[21] P. Kawa, M. Plata, and P. Syga. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. *arXiv preprint:2206.13979*, 2022. 8

[22] P. Kawa, M. Plata, and P. Syga. Defense against adversarial attacks on audio deepfake detection. In *IEEE Interspeech*, 2023. 6

[23] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56(Suppl 1):513–566, 2023. 1, 2

[24] P. Korshunov and S. Marcel. Improving generalization of deepfake detection with data farming and few-shot learning. *IEEE TBIOM*, 4(3):386–397, 2022. 8

[25] B. Kumar and S. R. Alraisi. Deepfakes audio detection techniques using deep convolutional neural network. In *IEEE COM-IT-CON*, volume 1, pages 463–468, 2022. 4

[26] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019. 2

[27] J. Li, C. Chen, M. R. Azghadi, H. Ghodosi, L. Pan, and J. Zhang. Security and privacy problems in voice assistant applications: A survey. *Computers & Security*, 134:103448, 2023. 2

[28] M. Li, Y. Ahmadiadli, and X.-P. Zhang. A comparative study on physical and perceptual features for deepfake audio detection. In *DDAM*, pages 35–41, 2022. 4

[29] M. Li, Y. Ahmadiadli, and X.-P. Zhang. A comparative study on physical and perceptual features for deepfake audio detection. In *DDAM*, page 35–41, 2022. 4

[30] X. Li, P.-Y. Chen, and W. Wei. Where are we in audio deepfake detection? a systematic analysis over generative and detection models. *ACM Transactions on Internet Technology*, 2025. 6

[31] Y. Li, M. Zhang, M. Ren, M. Ma, D. Wei, and H. Yang. Cross-domain audio deepfake detection: Dataset and analysis. *arXiv preprint arXiv:2404.04904*, 2024. 5

[32] G. Lin, W. Luo, D. Luo, and J. Huang. One-class neural network with directed statistics pooling for spoofing speech detection. *IEEE TIFS*, 19:2581–2593, 2024. 4

[33] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng. Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation. *arXiv preprint arXiv:2409.14743*, 2024. 5

[34] F. Lux, J. Koch, and N. T. Vu. Exact prosody cloning in zero-shot multispeaker text-to-speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 962–969, 2023. 2

[35] J. M. Mart'in-Donas and A. Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. *ICASSP*, pages 9241–9245, 2022. 4

[36] C. O. Mawalim, Y. Wang, A. Adila, S. Okada, and M. Unoki. Multilingual audio deepfakes dataset for robust and generalizable detection, 2025. 5

[37] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM MM*, page 2823–2832, 2020. 7

[38] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger. Does audio deepfake detection generalize? *Interspeech 2022*, 2022. 6, 8

[39] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv:2401.09512*, 2024. 5

[40] N. M. Müller, K. Pizzi, and J. Williams. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pages 85–91, 2022. 6, 7

[41] S. Muppalla, S. Jia, and S. Lyu. Integrating audio-visual features for multimodal deepfake detection. *arXiv preprint arXiv:2310.03827*, 2023. 4

[42] S. Muppalla, S. Jia, and S. Lyu. Integrating audio-visual features for multimodal deepfake detection. *arXiv preprint arXiv:2310.03827*, 2023. 8

[43] N. Naffi. Deepfakes: How to empower youth to fight the threat of misinformation and disinformation. http://tinyurl.com/ytmf8hse, 2024. 8

[44] A. Qureshi and D. Megias Jimenez. Blockchain-based multimedia content protection: Review and open challenges. *Applied Sciences*, 11(1):1, 2020. 7

[45] R. Ranjan, M. Vatsa, and R. Singh. Uncovering the deceptions: An analysis on audio spoofing detection and future prospects, 2023. 2

[46] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro. Timit-tts: A text-to-speech dataset for multimodal synthetic media detection. *IEEE access*, 11:50851–50866, 2023. 5

[47] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang. How good is chatgpt at audiovisual deepfake detection: A comparative study of chatgpt, ai models and human perception. *arXiv e-prints*, pages arXiv–2411, 2024. 7

[48] V. sharma. Hindustantimes, 2023. Accessed on July 18, 2023. 1

[49] sheeranalyticsandinsights. sheeranalyticsandinsights, 2022. Accessed on August, 2022. 1

[50] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *INTERSPEECH*, 2015. 7

[51] G. Ulutas, G. Tahaoglu, and B. Ustubioglu. Deepfake audio detection with vision transformer based method. In *IEEE TSP*, pages 244–247, 2023. 4

[52] P. van Rijn, S. Mertes, D. Schiller, P. Dura, H. Siuzdak, P. Harrison, E. André, and N. Jacoby. Voiceme: Personalized voice generation in tts. *arXiv preprint arXiv:2203.15379*, 2022. 3

[53] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 3

[54] C. Wang, J. Yi, J. Tao, H. Sun, X. Chen, Z. Tian, H. Ma, C. Fan, and R. Fu. Fully automated end-to-end fake audio detection. In *DDAM*, pages 27–33, 2022. 4

[55] X. Wang, H. Delgado, H. Tak, J.-W. Jung, H.-J. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, et al. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8. ISCA, 2024. 3, 5

[56] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, S. Zhang, and Z. Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv e-prints*, pages arXiv–2409, 2024. 3

[57] Y. Xie, H. Cheng, Y. Wang, and L. Ye. Domain generalization via aggregation and separation for audio deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2023. 4

[58] Y. Xie, Y. Lu, R. Fu, Z. Wen, Z. Wang, J. Tao, X. Qi, X. Wang, Y. Liu, H. Cheng, et al. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *arXiv preprint arXiv:2405.04880*, 2024. 5

[59] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu. Fakesound: Deepfake general audio detection. *arXiv preprint arXiv:2406.08052*, 2024. 5

[60] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *arXiv preprint arXiv:2208.05845*, 2022. 8

[61] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li. Audio deepfake detection system with neural stitching for add 2022. In *IEEE ICASSP*, pages 9226–9230, 2022. 4

[62] Z. Yan, Y. Zhao, and H. Wang. Voicewukong: Benchmarking deepfake voice detection. *arXiv preprint arXiv:2409.06348*, 2024. 5

[63] W. Yang, Y. Li, M. Fang, Y. Wei, T. Zhou, and L. Chen. Who can withstand chat-audio attacks? an evaluation benchmark for large language models. *arXiv preprint arXiv:2411.14842*, 2024. 1

[64] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE TIFS*, 18:2015–2029, 2023. 4, 5

[65] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang. A robust audio deepfake detection system via multi-view feature. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13131–13135. IEEE, 2024. 6

[66] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. DAI, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv e-prints*, pages arXiv–2502, 2025. 3

[67] Z. Yu, Y. Chang, N. Zhang, and C. Xiao. {SMACK}: Semantically meaningful adversarial audio attack. In *USENIX Security*, pages 3799–3816, 2023. 6

[68] C. Yushen, Z. Niu, Z. Ma, K. Deng, C. Wang, K. Yu, X. Chen, et al. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. 3

[69] X. Zhang, L. Xue, Y. Gu, Y. Wang, J. Li, H. He, C. Wang, S. Liu, X. Chen, J. Zhang, et al. Amphion: an open-source audio, music, and speech generation toolkit. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 879–884. IEEE, 2024. 3

[70] Y. Zhang, W. Lin, and J. Xu. Joint audio-visual attention with contrastive learning for more general deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):1–23, 2024. 4

[71] Z. Zhang, W. Hao, A. Sankoh, W. Lin, E. Mendiola-Ortiz, J. Yang, and C. Mao. I can hear you: Selective robust training for deepfake audio detection. *arXiv preprint arXiv:2411.00121*, 2024. 5

[72] Y. Zhu, S. Koppisetti, T. Tran, and G. Bharaj. Slim: Style-linguistics mismatch model for generalized audio deepfake detection. *arXiv preprint arXiv:2407.18517*, 2024. 4