Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Replay Attacks Against Audio Deepfake Detection

Nicolas Müller[1,3], *Piotr Kawa*[2,3], *Wei-Herng Choong*[1], *Adriana Stan*[4], *Aditya Tirumala Bukkapatnam*[3], *Karla Pizzi*[5,6], *Alexander Wagner*[1], *Philip Sperl*[1]

[1]Fraunhofer AISEC, Germany [2]Wrocław University of Science and Technology, Poland
[3]Resemble AI, USA [4]Technical University of Cluj-Napoca, Romania
[5]Neodyme AG, Germany [6]TU Munich, Germany,

nicolas.mueller@aisec.fraunhofer.de

Abstract

We show how replay attacks undermine audio deepfake detection: By playing and re-recording deepfake audio through various speakers and microphones, we make spoofed samples appear authentic to the detection model.

To study this phenomenon in more detail, we introduce *ReplayDF*, a dataset of recordings derived from M-AILABS and MLAAD, featuring 109 speaker-microphone combinations across six languages and four TTS models. It includes diverse acoustic conditions, some highly challenging for detection.

Our analysis of six open-source detection models across five datasets reveals significant vulnerability, with the top-performing W2V2-AASIST model's Equal Error Rate (EER) surging from 4.7% to 18.2%. Even with adaptive Room Impulse Response (RIR) retraining, performance remains compromised with an 11.0% EER. We release *ReplayDF* for non-commercial research use.

Index Terms: audio deepfake detection, anti-spoofing, dataset, replay attack, text-to-speech, voice cloning

1. Introduction

Text-to-speech (TTS) and Voice Conversion (VC) technologies have facilitated numerous advancements across various domains. In the entertainment industry, they enable independent studios in television and film to generate a diverse array of character voices, including the recreation of legacy character voices. In the healthcare sector, TTS has demonstrated significant potential in assisting individuals with speech impairments, as illustrated by initiatives such as Google's Parrotron [1]. Despite its potential, speech synthesis also presents significant risks, particularly through the misuse of replicating an unknowing target speaker's identity, also known as deepfake technology. These artificially generated identities can be exploited to spread disinformation, undermine trust in authorities, and harm individuals' reputations [2, 3]. Deepfakes have also been employed in fraudulent schemes, voice phishing attacks (vishing), and even in state-on-state warfare [4–6]. Addressing these challenges requires robust deepfake detection systems to determine whether an audio sample is genuine (*bona fide*) or fabricated (*spoofed*).

In the domain of speaker identification and authentication, three distinct attack scenarios emerge: a) *physical attacks*: used for access spoofing, this involves replay attacks targeting voice-biometric systems through the playback of recorded bona fide utterances; b) *logical access*: TTS and VC technologies are used to circumvent voice biometry systems by synthesizing a target speaker's voice characteristics; and c) *deepfake audio*: synthetic voice content designed to deceive human listeners [7], distributed primarily through social media platforms.

Figure 1: *Photographs from the recording labs, where we play bona fide and spoofed audio samples over a loudspeaker, and record them via microphone.*

In this context, we identify a critical gap: the vulnerability of deepfake detection systems to physical or replay attacks. Specifically, we examine scenarios where attackers attempt to compromise the detection system by re-recording the synthesised audio. Our analysis demonstrates that such replay attacks can successfully disguise audio deepfakes as genuine recordings, likely by removing subtle artifacts that detection models rely on for identification.

To address this research gap, we:

- introduce *ReplayDF*, a comprehensive dataset[1] of deepfake recordings. It comprises 109 combinations of loudspeaker and microphone across six languages and four text-to-speech (TTS) systems, plus Mean Opinion Scores;
- evaluate existing audio deepfake detection models using this dataset, demonstrating that *replay attacks* effectively disguise fake audio as genuine;
- show a correlation between model performance and recording quality; and
- provide evidence that adaptive retraining using room impulse responses (RIRs) from *ReplayDF* can help to lessen the effects of replay attacks.

[1]https://deepfake-total.com/replay_df

10.21437/Interspeech.2025-20

2. Related Work

Audio deepfake detection has been largely driven by ASVspoof [7–9], a challenge that initially focused on voice biometrics but recently expanded to include deepfake detection. The urgency of the problem is underscored by the rise of commercial text-to-speech providers like Resemble AI, Respeecher, and ElevenLabs [10–12], some of which have been involved in misinformation campaigns [13]. Research advancements in neural audio deepfake detection span both front-end and back-end components. Front-ends extract features from audio, evolving from time-frequency representations like mel-spectrograms [14] to raw waveforms (such as RawNet2 [15]) and self-supervised learning methods like Wav2Vec2 [16]. Back-ends focus on classifying the samples suing the extracted features, and exploring advanced neural architectures such as graph attention networks and conformers [17, 18]. Challenges in audio deepfake detection include generalization, which evaluates the discrepancy in performance on seen versus unseen data [19, 20], explainability of the model's decision [21, 22], as well as vulnerability to replay attacks.

Replay attacks were originally formulated in the ASVspoof challenge [8] in the context of "liveliness detection" to study the vulnerability of speaker authentication systems against recorded voice trials. The concept of recorded audio is also relevant in the context of adversarial evasion attacks, where the goal is to create an imperceptible perturbation δ such that a given model f misclassifies an input x, i.e. $f(x+\delta) \neq y$ such that δ is small w.r.t some norm. It has been shown that the creation of adversarial samples that survive the "air-gap" is especially challenging and requires specialized techniques [23–25]. This is because when presenting and capturing such an adversarial sample, the perturbation δ can be lost or altered in the process, leading to the deliberate manipulations of the original input to no longer be effective. In the context of replay attacks on audio deepfake detection, there is little related work. Luong et al. [26] demonstrate that RIRs degrade the performance of detection systems on ASVspoof 2021. However, their study is limited to simulated recordings rather than real-world replays.

This work investigates the replay attack phenomenon into greater depth by conducting controlled "air-gapping" experiments in a laboratory environment, where audio deepfakes are played back and re-recorded.

3. ReplayDF Database

To systematically assess the impact of replay attacks on deepfake detection, we introduce *ReplayDF*, a dataset of audio recordings generated by playing and re-recording both bona fide and spoofed samples using a diverse set of loudspeakers and microphones. The dataset spans six languages and incorporates attacks from four TTS models, ensuring an equal distribution between spoof and bona fide samples.

The data generation pipeline, outlined in Section 3, systematically selects audio samples from the MLAAD v5 dataset for spoofed recordings and the M-AILABS dataset for bona fide samples. Each selected sample is played through a loudspeaker and recorded using a microphone, thereby capturing playback distortions introduced by real-world acoustic environments and hardware characteristics. This procedure is repeated across multiple recording setups, resulting in a dataset encompassing 109 unique recording configurations and totaling 132.5 hours of audio data. Each audio sample is accompanied by complete metadata, as listed in Table 1: the original and recorded file paths, attack type (i.e., the TTS model if the original audio is spoofed), language, recording hardware, and setup images. A combination of loudspeaker and microphone is referred to as a *setup*, uniquely identified by a hash-based *uid*. To evaluate the quality of the dataset, we perform subjective listening tests and rate the quality of the recordings using Mean Opinion Scores (MOS), as detailed in Section 4.6.

Attribute	Info
original_file	Relative path to original file from either `MLAAD` or `M-AILABS`.
recorded_file	Relative path to recorded file.
label	`spoof` or `bona fide`.
architecture	`bark`, `vits`, `XTTS v1.1`, or `XTTS v2.0`.
language	`en`, `de`, `fr`, `it`, `pl`, or `sp`.
mic	Description of the microphone used.
speaker	Description of the loudspeaker used.
uid	Unique folder identifier.
setup.jpg	Photograph of the recording setup.
RIR.wav	Room Impulse Response of recording setup.

Table 1: *Metadata for* ReplayDF.

Algorithm 1 The data creation pipeline for *ReplayDF*. For each setup (i.e., per combination of loudspeaker and microphone), we select $n = 10$ instances for each language and TTS model from both MLAAD v5 (spoof) and M-AILABS (bona fide). Recordings and original audio files are stored to create a balanced dataset of air-gapped vs. non-air-gapped data.

$I = [id_0, id_1, ..., id_{108}]$ ▷ Speaker/microphone used.
$n = 10$
$R = \{\}$ ▷ List to hold recorded audio files.
$O = \{\}$ ▷ List to hold original audio files.
for id ∈ I **do**
  for lang ∈ {en, de, fr, it, pl, es} **do**
    for model ∈ {bark, vits, xtts_v1.1, xtts_v2.0} **do**
      $A \leftarrow$ choose n from `M-AILABS(lang)`
      $B \leftarrow$ choose n from `MLAAD(lang, model)`
      for w ∈ $A \cup B$ **do**
        r = `play_and_record(id, w)`
        $R \leftarrow R \cup \{r\}$
        $O \leftarrow O \cup \{s\}$
      end for
    end for
  end for ▷ $6 \cdot 4 \cdot 2 \cdot 10 = 480$ recordings per id.
end for ▷ Total number of recordings: $|I| \cdot 480 = 52320$.

4. Experiments

4.1. Evaluation Approach

We evaluate *ReplayDF* across multiple scenarios to assess the impact of replay attacks on audio deepfake detection models. We define two key data partitions: first, *ReplayDF* (set R): All audio files generated as in Section 3, containing equal amounts of bona fide and spoofed instances. Second, the *Baseline* dataset (set O): the original input instances from MLAAD v5 and M-AILABS, serving as a comparative baseline against *ReplayDF*.

Model	Accuracy (%) ↑		EER (%) ↓	
	Baseline	ReplayDF	Baseline	ReplayDF
Whisper [27]	57.9	50.0	44.7	49.5
Raw PC Darts [28]	69.4	56.6	32.1	43.9
RawNet2 [15]	74.3	57.1	25.9	43.1
TCM ADD [18]	73.5	59.6	13.3	37.3
RawGAT-ST [29]	79.4	58.7	19.8	40.2
W2V2-AASIST [30]	90.0	74.2	10.6	24.8

Table 2: *Performance of Open-Source models with publicly available checkpoints in mean accuracy and EER over ReplayDF, as well as the original audio files (Baseline). The threshold for accuracy computation is as specified in the respective original publication. In all scenarios, replay attacks deteriorate model performance.*

Training Dataset	Accuracy (%) ↑		EER (%) ↓	
	Baseline	ReplayDF	Baseline	ReplayDF
ASVspoof 19	74.6 ± 6.6	55.1 ± 3.2	14.3 ± 7.7	34.7 ± 3.1
ASVspoof 5	67.6 ± 8.6	52.8 ± 1.4	12.5 ± 4.1	34.8 ± 3.5
Fake-or-Real	54.2 ± 0.9	49.5 ± 0.0	28.4 ± 1.6	45.0 ± 1.3
In-the-Wild	66.8 ± 3.1	52.9 ± 1.4	30.2 ± 3.2	42.4 ± 1.2
ODSS	94.7 ± 0.7	77.7 ± 2.4	4.7 ± 0.9	18.2 ± 1.5

Table 3: *Performance of W2V2-AASIST, trained on five different datasets, and evaluated on* ReplayDF. *Results computed over three independent trials, with mean and standard deviation shown.*

4.2. Publicly Available Deepfake Detection Checkpoints

The first evaluation we perform with *ReplayDF* is against open-source deepfake detection models. We use publicly available checkpoints and their original hyperparameters and rate their performace for the O and R subsets of *ReplayDF*. Table 2 summarizes the results, highlighting performance differences in percentage points between the two subsets. Across all tested models, we observe a significant performance degradation of up to 20 percentage points when transitioning from *Baseline* to *ReplayDF*.

4.3. Dataset Selection

To evaluate this phenomenon in more detail, we chose the best-performing model architecture from Table 2, i.e. W2V2-AASIST, and re-train it on five different audio deepfake datasets: ASVspoof2019 [8], ASVspoof 5 [9], Fake-or-Real [31], In-the-Wild [19] and the Open Dataset of Synthetic Speech (ODSS) [32]. The models are trained for 75 epochs using early stopping based on the training loss, employing the Adam optimizer with a learning rate of $4 \cdot 10^{-6}$ and a batch size of 32.

We then evaluate the retrained models on *ReplayDF* and the *Baseline* dataset. Table 3 presents the results in terms of mean and standard deviation of the accuracy and EER, over three independent trials. Given that the ODSS-trained model performs best, we select this configuration for the following evaluations. Note that this experiment re-confirms the results from Table 2: irrespective of training dataset, the model performance deteriorates when moving from the *Baseline* dataset to *ReplayDF*. Even in the case of the best model trained on ODSS, the deterioration is significant, dropping from 4.7% to 18.2% EER.

Attack	No Augmentation		With Augmentation	
	Baseline	ReplayDF	Baseline	ReplayDF
Bark	82.6	40.7	83.0	56.9
VITS	82.2	53.4	75.7	65.8
XTTS v1.1	100.0	66.3	99.7	76.2
XTTS v2	100.0	59.4	99.8	73.6
bona fide	98.3	97.7	99.9	98.6

Table 4: *Detection accuracy [%] (↑) of W2V2-AASIST on the* Baseline *and* ReplayDF*, with and without RIR augmentation. While TTS-generated spoofs cause substantial accuracy reductions (up to -43.6 percentage points), bona fide detection remains unaffected. Incorporating RIR augmentation during training reduces the effect of the replay attacks.*

4.4. Analysis of False Positives and Negatives

An essential analysis is to evaluate whether the overall decrease in model performance is caused by an increased number of false positives or false negatives. In other words: are bona fide instances mistakenly classified as spoof (false positives) or are the spoofed instances missed by the model (false negatives)? Table 4 details the results obtained by W2V2-AASIST model trained on ODSS. It shows the model accuracy[2] for each of the individual attacks in *ReplayDF* (Bark, VITS, XTTS v1.1 and v2.0) and for the bona fide instances. It can be clearly observed that the air-gap affects solely the spoofed instances, which exhibit a decrease in performance between 28 and 42 absolute percent points, while the performance over the bona fide instances remains unchanged.

Finally, we note that models trained on datasets where noise patterns differ between genuine and spoofed audio (like older datasets such as In-the-Wild or ASVspoof 19) can develop problematic shortcuts [21, 33]: they learn to equate poor audio quality with spoofed samples. This leads to a misclassification of the entire *ReplayDF* data as spoofed audio, simply because recordings (both bona fide and spoof) contain channel noise.

4.5. Adaptive Defender

Next, we evaluate if augmenting the training data with room impulse responses (RIRs) from the dataset can help defend against replay attacks. During training, we convolve the training data with RIRs from *ReplayDF*, then evaluate performance on both the *Baseline* and *ReplayDF* datasets as before.

RIR augmentation reduces the replay attack's impact, improving the overall EER from 18.2% to 11.09% (compared to the baseline EER of 4.7% and 2.2% without and with RIR augmentation, respectively). The right-hand side of Table 4 shows results per attack type. While *Baseline* performance remains stable, accuracy on *ReplayDF* improves by about 10 to 15 percentage points for each of the attacks. Thus, while RIR augmentation during training does improve resilience to replay attacks, it falls short of fully mitigating this vulnerability.

[2]Since we compute the performance per attack, we only have a single label and cannot compute EER. Instead, we report accuracy, where predictions are classified as positive if they exceed a threshold of 50%. Notably, since *ReplayDF* is perfectly balanced, accuracy serves as an appropriate metric in this context.

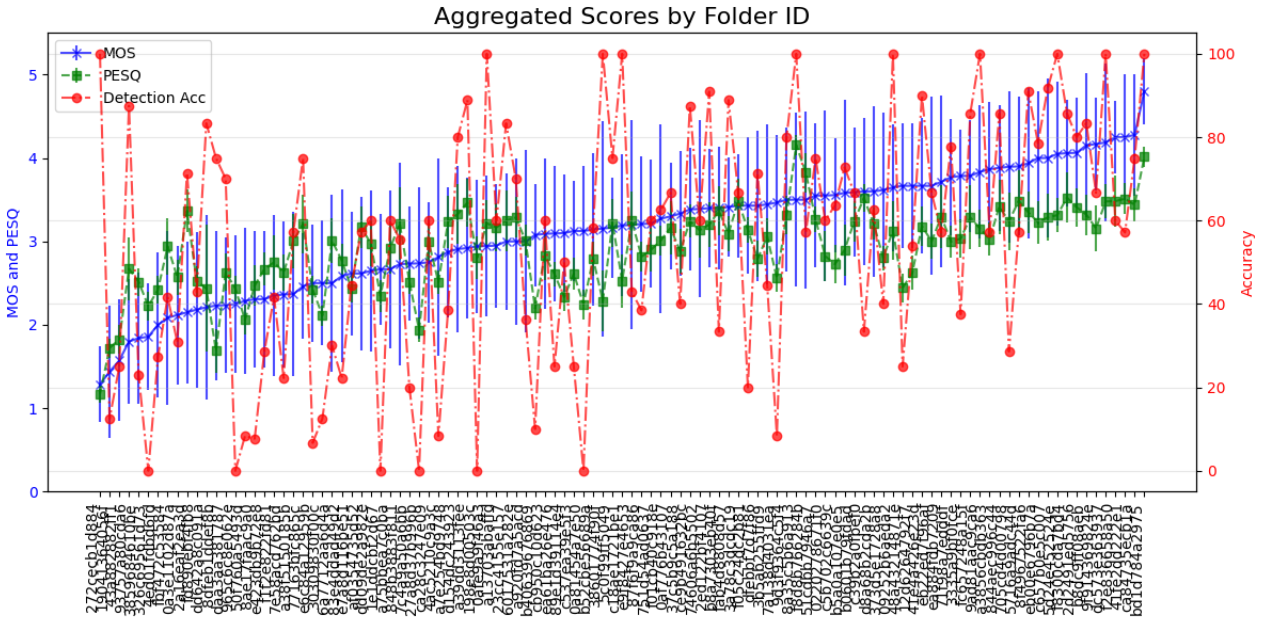

Figure 2: *Overview of recording quality in* ReplayDF*, measured by MOS and PESQ (blue, green). Detection performance on audio deepfakes (red) correlates with recording quality, showing Pearson correlations of 0.423 and 0.509, respectively. This suggests that the more aggressive the replay attack, the worse the detection performance.*

4.6. Recording Quality vs. Detection Performance

To further investigate the models' performance deterioration, we analyze whether it correlates with potential quality degradation in the recordings. This analysis is conducted using two measures. First, we use the Perceptual Evaluation of Speech Quality (PESQ) [34] score, which is an objective metric used to assess audio quality by comparing a reference and degraded signal. It provides a score ranging from -0.5 to 4.5, where higher values indicate better quality. Second, we perform human listening tests, where subjects are asked to score bona fide samples from *ReplayDF* on a score between 1 (worst) and 5 (best). We obtain a Mean Opinion Score (MOS) for each setup UID by averaging three scores from at least four listeners per UID.

Figure 2 displays recording UIDs against PESQ and MOS, as well as the accuracy of the W2V2-AASIST deepfake detection system. The analysis reveals that our database encompasses a wide range of recording quality levels, with MOS and PESQ values varying from low (i.e. < 2) to excellent (> 4). Furthermore, MOS and PESQ scores correlate with a Pearson coefficient of 0.681, confirming consistency between subjective and objective quality assessments. The deepfake detection accuracy exhibits a Pearson correlation of 0.423 with MOS and 0.509 with PESQ. This suggests that lower-quality replay attacks reduce the detection effectiveness of audio deepfake systems.

4.7. Recordings vs. Noise

Finally, we investigate whether the performance drop in deepfake detection is primarily caused by general quality degradation, such as added noise, or rather by the loss of distinctive model- or deepfake-specific characteristics due to the air-gap. To address this, we add noise to the *Baseline* dataset and evaluate the previously trained model on it. We add three different types of noise (Gaussian, white, and pink noise) at varying signal-to-noise ratios (SNR). For each input audio file, a synthetic noise of matching length is generated and mixed with the original audio at a randomly selected target SNR level between 15-40 dB (the SNR value is recorded in the metadata). To achieve the target SNR, we estimate the signal power and generate a noise sample of equal number of samples. Given the noise signal's power and a target SNR value in dB, a linear scaling factor is calculated scale the noise signal down, such that the mixture of the audio and noise signals is at the desired SNR.

We find that model performance is not impacted much. W2V2-AASIST exhibits accuracy changes of -2.3% (Bark), -0.2% (XTTS v2), $+0.0\%$ (XTTS v1.1), and $+3.1\%$ (VITS), while bona fide detection remains unaffected ($+0.0\%$). These findings suggest that the performance degradation observed in replay attack scenarios is not merely due to noise. Given that Luong et al. [26] report similar findings on *simulated* recordings, improving deepfake detection models to better handle convolutional noise—such as distortions introduced by room impulse responses—could be a crucial step toward enhancing their resilience against attacks.

5. Conclusion

In this work, we investigate the impact of replay attacks on audio deepfake detection systems by introducing *ReplayDF*, a comprehensive dataset of 132.5 hours of re-recorded spoof and bona fide audio. Our results demonstrate that replay attacks significantly degrade detection performance, effectively disguising deepfake audio as authentic, while bona fide samples remain unaffected. We further show that simply adding noise does not lead to the same degradation, indicating that replaying removes key artifacts relied upon by detection models. To support future research in this area, we publicly release *ReplayDF* along for non-commercial use.

Acknowledgement. This work was partially funded by project DLT-AI SECSPP (id: PN-IV-P6-6.3-SOL-2024-2-0312) and the Department of Artificial Intelligence, Wrocław University of Science and Technology.

6. References

[1] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech 2019*, 2019, pp. 4115–4119.

[2] "How deepfake videos are used to spread disinformation - the new york times," https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html, (Accessed: 16.10.2024).

[3] K. Tenbarge, "Taylor swift nude deepfake goes viral on x, despite platform rules," 1 2024, [Online; accessed 03.02.2025]. [Online]. Available: https://www.nbcnews.com/tech/misinformation/taylor-swift-nude-deepfake-goes-viral-x-platform-rules-rcna135669

[4] "A voice deepfake was used to scam a ceo out of $243,000," https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/, (Accessed: 16.10.2024).

[5] "NSE CEO deepfake: NSE urges caution after fake videos of CEO Ashish Chauhan recommending stocks go viral - The Economic Times," https://economictimes.indiatimes.com/markets/stocks/news/beware-of-deepfake-of-ceo-recommending-stocks-says-nse/articleshow/109189329.cms, (Accessed: 16.10.2024).

[6] "A deepfake video showing volodymyr zelenskyy surrendering worries experts : Npr," https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia, (Accessed: 16.10.2024).

[7] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. of Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[9] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *ASVspoof Workshop 2024*, 2024.

[10] Resemble AI, "AI Voice Cloning: Clone your Voice in Seconds," https://www.resemble.ai/voice-cloning/, 2024, Accessed: 17.10.2024.

[11] Respeecher, "AI Voice Cloning," https://www.respeecher.com/ai-voice-cloning, 2024, Accessed: 17.10.2024.

[12] Eleven Labs, "Create a replica of your voice that sounds just like you," https://elevenlabs.io/voice-cloning, 2024, Accessed: 17.10.2024.

[13] K. Knibbs, "Researchers say the deepfake biden robocall was likely made with tools from ai startup elevenlabs — wired," 1 2024, [Online; accessed 2025-01-14]. [Online]. Available: https://www.wired.com/story/biden-robocall-deepfake-elevenlabs/

[14] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Interspeech 2021*, 2021, pp. 4259–4263.

[15] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-End anti-spoofing with RawNet2," in *IEEE ICASSP 2021*, 2021, pp. 6369–6373.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[17] J. Jung, H. Heo, H. Tak, H. Shim, J. Chung, B. Lee, H. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. of ICASSP*, 2022.

[18] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," *arXiv preprint arXiv:2406.17376*, 2024.

[19] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Interspeech*, 2022.

[20] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, "Towards generalisable and calibrated audio deepfake detection with self-supervised representations," in *Proc. of Interspeech*, 2024.

[21] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?" in *Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[22] G. Channing, J. Sock, R. Clark, P. Torr, and C. S. de Witt, "Toward robust real-world audio deepfake detection: Closing the explainability gap," *arXiv preprint arXiv:2410.07436*, 2024.

[23] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "Adversarial attack and defense: A survey," *Electronics*, vol. 11, no. 8, p. 1283, 2022.

[24] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc of ICML*, 2019.

[25] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. of ICML*, 2018.

[26] H.-T. Luong, D.-T. Truong, K. A. Lee, and E. S. Chng, "Room impulse responses help attackers to evade deep fake detection," in *2024 IEEE SLT Workshop*. IEEE, 2024, pp. 623–629.

[27] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 4009–4013.

[28] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," in *Proc. of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 22–28.

[29] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[30] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. of Speaker and Language Recognition Workshop*, 2022.

[31] "Datasets – APTLY and LaSSoftE," [Accessed 24.01.2025]. [Online]. Available: https://bil.eecs.yorku.ca/datasets/

[32] A. Yaroshchuk, C. Papastergiopoulos, L. Cuccovillo, P. Aichroth, K. Votis, and D. Tzovaras, "An Open Dataset of Synthetic Speech," in *IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2023, pp. 1–6.

[33] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, 2020.

[34] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP*, 2001.