Generated Audio Detectors are Not Robust in Real-World Conditions

Soumya Shaw [1] **Ben Nassi** [2] **Lea Schönherr** [1]

Abstract

The misuse of generative AI (genAI) has raised significant ethical and trust issues. To mitigate this, substantial focus has been placed on detecting generated media, including fake audio. In this paper, we examine the efficacy of state-of-the-art fake audio detection methods under real-world conditions. By analyzing typical audio alterations of transmission pipelines, we identify several vulnerabilities: (1) minimal changes such as sound level variations can bias detection performance, (2) inevitable physical effects such as background noise lead to classifier failures, (3) classifiers struggle to generalize across different datasets, and (4) network degradation affects the overall detection performance. Our results indicate that existing detectors have major issues in differentiating between real and fake audio in practical applications and that significant improvements are still necessary for reliable detection in real-world environments.

1. Introduction

The innovative use of generative AI (genAI) is truly astounding, having shifted the spotlight from the industrial era to the digital age, revolutionizing every facet of how we live, work, and interact with technology. Yet, genAI came under scrutiny once more as the world observed that the technology can also be used to delve into deceptive practices: Generated media have been used maliciously to manipulate voters (Seitz-Wald & Memoli, 2024) and to fraud companies (Chen & Magramo, 2024). With the power to replicate and generate content with astonishing precision, it raises concerns about the decline in trust, the spread of disinformation, and the ethical implications surrounding the use of this technology (de Ruiter, 2021). While generated images, videos, and text have garnered significant attention (Wang et al., 2020; Li & Lyu, 2019; Mitchell et al., 2023) over the years, another aspect of this phenomenon has also begun to capture the limelight: generated audio and speech. Generated audio has become a popular tool for scammers due to its broad reach and critical role in the creation of misleading videos.

Securing against such scamming attempts using generated audio has been an open problem, picking up pace in recent times. In that regard, several fake audio detectors have been proposed (Ballesteros et al., 2021; Zhang et al., 2021; Ge et al., 2021; Tak et al., 2021), to automatically detect generated media. Although these detectors yield remarkable results on unaltered generated audio, their robustness in real world settings remains questionable.

In this paper, we investigate state-of-the-art fake audio detection methods under real-world conditions. For this, we analyze audio alterations from the perspective of the audio transmission pipeline, such as communication channels and transmissions. We show that (1) minimal changes, such as sound level differences can significantly compromise their performance with a bias, (2) physical effects (e.g., background noise) that are typically inevitable for transmission can also cause the classifiers to fail, (3) the classifiers do not generalize to other datasets, and (4) network degradation impact their overall performance. In general, classifiers tend to classify every sample as fake or real, regardless of the origin. This effect is particularly present for changing sound levels, where we can mostly control the classification of a sample with it sound level. Our experiments indicate that current detectors cannot reliably discern real and fake audio samples in real-world settings.

In summary, we make the following key contributions: (1) **Audio Transmission Pipeline.** We investigate different alterations to the audio on amplitude level, as well as other degradations for digital transmission, and demonstrate their impact on recognition performance. (2) **Out-of-Distribution (OOD) Samples.** We test the performance of state-of-the-art classifiers on recent audio generation models and establish that the tested classifiers cannot robustly distinguish real and fake audio for OOD samples.

[1] CISPA – Helmholtz Center for Information Security, Saarbrücken, Germany [2] Cornell Tech, New York, USA. Correspondence to: Soumya Shaw <soumya.shaw@cispa.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

2. Background

Synthetically-Generated Audios. Synthetic audio generators utilize machine learning algorithms, particularly based on deep neural networks, to generate and replicate human voices convincingly. Traditional algorithms analyze vast amounts of audio data to understand the intricacies of speech patterns, vocal tone, and even specific nuances unique to an individual's voice (Lukose & Upadhya, 2017; Tabet & Boughazi, 2021). Recent algorithms even surpass their predecessors by implementing one-shot or few-shot architectures, requiring only a few seconds of voice to be able to clone a voice (Huang et al., 2022; Wu et al., 2022; Xue et al., 2022) while achieving real-time voice cloning (Jemine, 2019). By harnessing this understanding, they can generate entirely new audio clips that sound eerily similar to the target speaker.

Detecting Synthetically-Generated Audios. To curb the misuse of generative technologies in audio, various deep neural network (DNN)-based detection methods have been developed. Fake audio detection is a binary classification, and its goal is to distinguish between fake and authentic audio recordings. The classifier is a function $f : \mathcal{A} \to \mathbb{R}$ that takes an audio sample $a \in \mathcal{A}$ as input and produces a real-valued scalar as output. A lower output value signifies a higher probability that the input audio is fake, and a higher value corresponds to a higher likelihood that the audio is real.

Audio Communication Pipeline. For the detection of generated audio, it is not enough to consider only the original generated audios. Audio is never used in isolation but is always transmitted via some kind of communication pipeline. Therefore, to effectively detect generated audio, we need to take into account the transmission pipeline.

Despite efforts to improve the performance of detectors for generated audio, for example, by augmenting the training data with modified versions of the original samples, these inadvertent changes can influence the performance of automatic detection methods. For instance, background noise from the surrounding environment can affect the distribution of the audio. Similarly, distortions in the input audio due to factors such as codecs, lossy compression, and network degradation provide room for additional distribution shifts.

3. Fake Audio Detectors and Datasets

In this section, we explain the fake audio detectors we analyzed and the datasets used in this study.

3.1. Fake Audio Detectors

We settle on four detection methods, based on their performance, publication year, and code availability:

- **ResNet-18:** The first model is based on ResNet-18 (Zhang et al., 2021) and uses the ASVspoof2019 dataset for training and testing (evaluation). The test accuracy is 0.945 and its AUC score is 0.995.
- **RawPC-DARTS:** The detector (Ge et al., 2021) also operates on the ASVspoof2019 dataset for training and testing. The test accuracy is 0.963 and the AUC Score is 0.995.
- **RawNet-2:** RawNet-2 (Tak et al., 2021) also utilizes the ASVspoof2019 dataset for training and testing. The model was trained by us since the pre-trained model is not provided. The model yields a test accuracy of 0.926 and an AUC score of 0.984.
- **Whisper Features:** The detector (Kawa et al., 2023) is trained on 125,000 samples of ASVspoof2021 dataset (Yamagishi et al., 2021) and 31,779 samples of DeepFakes In-The-Wild dataset (Müller et al., 2022). The MFCC joint features for MesoNet architecture was selected among the options, which was concluded the best performing combination in their paper. The test accuracy was found to be 0.942 and AUC score to be 0.973.

As is evident, the fake audio detectors mentioned above produce high accuracy (> 0.92) and AUC scores (> 0.97) in the ASVspoof2019 dataset.

3.2. Datasets

We conduct an extensive literature review to unravel the datasets used by fake audio detection models since 2015 and found that a staggering 84% of papers rely on one of the ASVspoof family datasets[1]. In addition, we collect an additional dataset consisting of several state-of-the-art generative models.

ASVspoof2021. The dataset consists of 611,829 audio samples with an average length of 2.99 seconds. The dataset was originally compiled to serve the task of synthetic audio detection and includes bonafide and spoofed utterances from various speakers and synthetic sources.

TTS Dataset. We compile another dataset, sourcing audio samples generated by current state-of-the-art TTS models

[1] The ASVspoof datasets have been released for the "Automatic Speaker Verification Spoofing And Countermeasures Challenges" since 2015. In total, there exist four different versions of the dataset, which also have overlapping subsets.

from both commercial and academic backgrounds. Amazon Polly (Amazon.com, Inc., 2023), GoogleTTS (Google, LLC, 2023), and ElevenLabs (ElevenLabs, Inc., 2023) are commercial TTS tools that we utilized in our data collection process. These models are closed-source, and there is no information available on their architecture and training data. Among the open-source models we selected, VITS (Kim et al., 2021), LST-TTS (Chen & Rudnicky, 2022), FastSpeech 2 (Ren et al., 2021), and, DiffGAN-TTS (Liu et al., 2022). VITS integrates variational inference with normalizing flows and adversarial training to generate more natural-sounding speech. LST-TTS is a transformer-based text-to-speech synthesis system that achieves fine-grained style control by using local style tokens and cross-attention blocks to fuse content and style information. In FastSpeech 2, the authors utilize high-performance alignment and pitch extraction tools for achieving high-quality and rapid speech synthesis. DiffGAN-TTS adopts an expressive model as a denoising function to approximate the true denoising distribution with adversarial training.

The dataset, which we will refer to as the "OOD dataset" for the rest of the paper, comprises 3024 validated audio samples from Common Voice Delta Segment 15.0 (Mozilla Foundation, and Community, 2017) classified as bonafide samples and an additional 3024 audio samples generated by seven TTS algorithms in total. Among these, 450 audio samples were synthesized by each generator, except for DiffGAN-TTS, ElevenLabsTTS, and GoogleTTS, which contributed 421, 333, and 470 samples respectively. The TTS algorithms utter the same content (text) from the bonafide segment to avoid any bias that may arise due to the content. The choice of TTS models was made primarily to have a fair distribution across different state-of-the-art architectures and publicly and commercially available tools.

4. Evaluation

In this section, we assess the effectiveness of fake audio detectors in real-world environments.

4.1. Evaluating a Baseline

We evaluate the performance of the chosen detectors on the ASVspoof2021 dataset, as shown in Table 1. The models perform extremely well on the dataset in terms of accuracy and show decent AUC scores. However, a degraded performance can be witnessed when the false positive rate (FPR) is fixed at 5%. The models seem to perform very well in identifying the fake audio but struggle with the real speakers.

Model Name	Accuracy	TPR@ 0.05 FPR	AUC Score
ResNet-18	0.942	0.522	0.756
RawPC-DARTS	0.973	0.623	0.798
RawNet-2	0.955	0.633	0.854
Whisper features	0.936	0.565	0.924

Table 1. Performance of selected detection models on the ASVspoof2021 dataset.

Model Name	Accuracy	TPR@ 0.05 FPR	AUC Score
ResNet-18	0.727	0.517	0.754
RawPC-DARTS	0.796	0.630	0.802
RawNet-2	0.787	0.634	0.854
Whisper features	0.819	0.576	0.924

Table 2. Performance of selected detection models on the balanced ASVspoof2021 dataset.

4.2. Balanced vs. Imbalanced Evaluation

We observe that the ASVspoof2021 dataset is highly unbalanced. Out of 611,829 audio samples from the eval set, 589,212 samples are generated by spoofed sources and only 22,617 audio samples are from authentic speakers. Using the biased dataset may misinterpret the detection model's performance on a real-world analysis, and hence arises the need to use an unbiased subset of the main dataset. To obtain such a subset, we randomly sample 20,000 audio samples from the original dataset, comprising 10,000 files from each class, and, in turn, term it the 'balanced dataset.' The performance on the balanced dataset is recorded in Table 2. We notice that the accuracy falls approximately by 15% due to the balancing, but the other metrics do not change. The balanced dataset thus appears to be a good subset for carrying out the experiments.

4.3. Generalization to OOD Dataset

Our next evaluation focuses on the performance of the detectors for the OOD dataset. The performance of the detectors on this data is shown in Table 3. We observe a significant drop for all three metrics compared to the balanced ASVspoof2021 dataset. Whisper features appears to be the detector that suffered the most performance degradation. The exception is ResNet-18 which gets better in classifying fake audios as the TPR is higher for the 0.05 FPR thresholds. However, the low accuracy and high AUC score of ResNet-18 indicate a strict imbalance in the classification result and that most of the real audio is also classified as fake.

Insight 1: *Fake audio detectors do not generalize to data from unseen generative models, marking a decrease in performance if generative models are used that are not part of the training set.*

Model Name	Accuracy	TPR@ 0.05 FPR	AUC Score
ResNet-18	0.595 (0.727)	0.860 (0.517)	0.952 (0.754)
RawPC-DARTS	0.642 (0.796)	0.320 (0.630)	0.686 (0.802)
RawNet-2	0.598 (0.787)	0.220 (0.634)	0.813 (0.854)
Whisper features	0.411 (0.819)	0.036 (0.576)	0.382 (0.924)

Table 3. Performance of selected detection models on the OOD dataset (comparison to balanced ASVspoof2021 in brackets).

Dataset	Model Name	Accuracy	TPR@ 0.05 FPR	AUC Score
Balanced ASVspoof2021	ResNet-18	0.715 (0.727)	0.503 (0.517)	0.740 (0.754)
	RawPC-DARTS	0.721 (0.796)	0.498 (0.630)	0.784 (0.802)
	RawNet-2	0.732 (0.787)	0.417 (0.634)	0.781 (0.854)
	Whisper features	0.797 (0.819)	0.417 (0.576)	0.873 (0.924)
Out-of-Distribution	ResNet-18	0.525 (0.595)	0.698 (0.860)	0.946 (0.952)
	RawPC-DARTS	0.508 (0.642)	0.334 (0.320)	0.649 (0.686)
	RawNet-2	0.536 (0.598)	0.062 (0.220)	0.483 (0.813)
	Whisper features	0.427 (0.411)	0.028 (0.036)	0.375 (0.382)

Table 4. Performance of selected detection models after 15 dB reduction in amplitude (Comparison to balanced ASVspoof2021 and OOD dataset in brackets respectively).

4.4. Evaluating the Influence of the Amplitude

We proceed to analyze the impact of input amplitude/power on fake audio detectors. It is important to note that in real-world scenarios, the amplitude/power of an audio input can vary due to differences in speech patterns, microphone quality, and the distance between the speaker and the microphone. A fake detector should be invariant of this factor and work robustly for different sound levels. In the first experiment, we reduce the amplitude of the target audio samples by 15 dB and evaluate the detectors on the resulting dataset.

We compare the results in Table 4 with the results from Table 2 and 3 in brackets. We observe a performance degradation for almost all measured metrics and models and observe a decrease in AUC score and TPR @ 0.05 FPR to near-random true positive rate (TPR) for the balanced ASVspoof dataset, pointing towards a shift in trend to classify every sample as real. For the OOD dataset, we also observe lower performance, except that RawNet-2 shows an extreme classification bias. Almost all samples are classified as real irrespective of the ground truth, which explains the slightly better accuracy (close to 50%) but the low AUC value. Whisper features' performance remains comparable to the previous results.

For a more in-depth evaluation, we also perform experiments where we modify the changing sound level between -1 dB and -20 dB. The results are shown in Figure 1 for the balanced ASVspoof2021 dataset. Here we plot the TPs and TNs for RawPC-DARTS and RawNet-2, which show the variation in the metrics regarding amplitude changes. Interestingly, we observe that the classification of TPs and TNs seems to be highly correlated with the amplitude level. In other words, everything below approximately -15 dB is more likely classified as real, while everything above this threshold is more likely classified as fake, regardless of whether the sample is originally real or fake, indicating that the models have a strong bias to the sound level. We observed a similar trend for the OOD dataset in Appendix A for Figure 5.

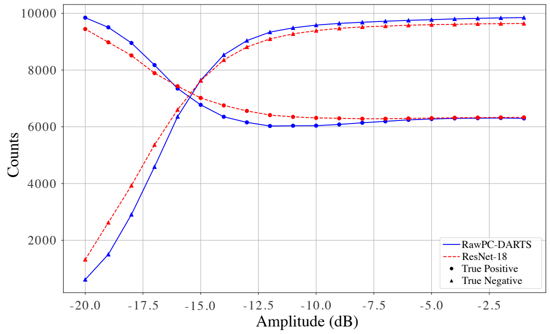

Figure 1. TPs and TNs for different amplitude levels over the balanced ASVspoof2021 dataset for RawPC-DARTS and RawNet-2.

> **Insight 2**: *Generated audio detectors show reduced performance when faced with amplitude changes in the input data and have a strong bias related to the sound level.*

4.5. Evaluating the Influence of the Background Noise

We then analyze the influence of background noise on the performance of the detectors. In reality, background noise can be introduced into an audio input due to factors such as wind, music, or crowded environments, and detectors should therefore be robust to changing conditions.

We design experiments to study the performance change induced by ambient noise. For this, we pick an ambient noise and iterate the experiment for varying signal-to-noise ratios (SNRs), namely 5 dB, 10 dB, 15 dB, and 20 dB, to closely model the effects of different noisy backgrounds in our study. The ambient noise is that of a bus with people chatting from *torchaudio*.

Figures 2, 3 and 4 show the results for the four tested models in terms of accuracy, TPR @0.05FPR and ROC AUC score respectively. The strength of the noise affects the classification capability of the models. However, it is particularly interesting to note how RawPC-DARTS and RawNet-2 break down completely for all three metrics at higher noise levels (lower SNRs). The sudden shift in results from 15 dB SNR to 10 dB SNR shows the migration of classification towards the negative class, i.e. more audio gets classified as fake,

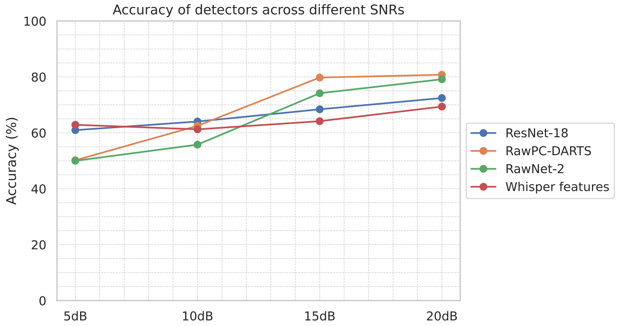

Figure 2. Accuracy of selected detection models on the balanced dataset.

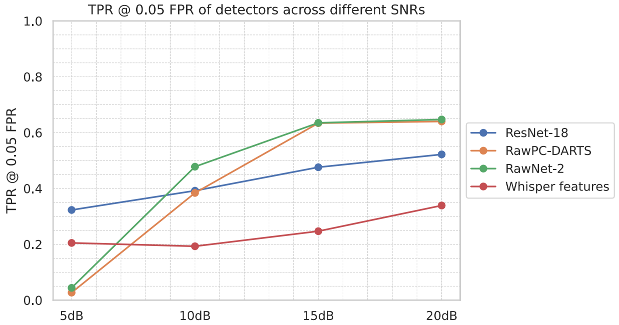

Figure 3. TPR @0.05 FPR of selected detection models on the balanced dataset.

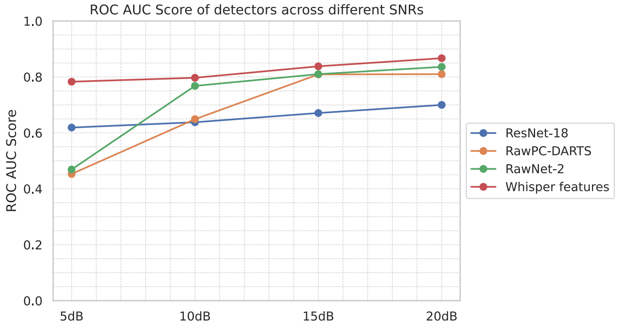

Figure 4. ROC AUC score of selected detection models on the balanced dataset.

Dataset	Model Name	Accuracy	TPR@ 0.05 FPR	AUC Score
Balanced ASVspoof2021	ResNet-18	0.546 (0.727)	0.054 (0.517)	0.574 (0.754)
	RawPC-DARTS	0.808 (0.796)	0.641 (0.630)	0.806 (0.802)
	RawNet-2	0.566 (0.787)	0.190 (0.634)	0.496 (0.854)
	Whisper features	0.828 (0.819)	0.548 (0.576)	0.913 (0.924)
Out-of-Distribution	ResNet-18	0.512 (0.595)	0.408 (0.860)	0.864 (0.952)
	RawPC-DARTS	0.642 (0.642)	0.320 (0.320)	0.685 (0.686)
	RawNet-2	0.489 (0.598)	0.078 (0.220)	0.534 (0.813)
	Whisper features	0.421 (0.411)	0.038 (0.036)	0.390 (0.382)

Table 5. Performance of selected detection models after applying Opus codec (Comparison to balanced ASVspoof2021 and OOD dataset in brackets respectively).

irrespective of its nature. Whisper features doesn't show drastic drops and performs similar even under increased noise levels. ResNet-18 also shows a smaller performance drop with increasing noise. However, the overall performance is already worse in comparison to the other models for higher SNRs.

Insight 3: *The performance of fake audio detectors decreases in response to noisy inputs, marking their limitation in handling audio in noisy environments.*

4.6. Codec Losses

Codecs are special software and hardware designed to compress and decompress digital audio signals for transmission. They are an integral part of the channel that ensures efficient transmission because of its reduction in storage size while maintaining acceptable levels of audio quality. We test the effects of one of the most common codecs used in recent times, namely Opus. Opus is an open, versatile, and highly efficient audio codec designed for interactive real-time applications such as VoIP, video conferencing, and online gaming, as well as for streaming and storage, making it one of the most widely used codecs (Valin et al., 2012; Xiph.Org, Foundation, 2012).

To formalize the experiment, we encode and decode audio samples with Opus and evaluate the resultant audio signals on the fake audio detection models. The results of the Opus codec are summarized in Table 5. ResNet-18 and RawNet-2 remain poor, especially in the balanced ASVspoof2021 dataset. Although similar trends can be found on the OOD dataset as well, RawNet-2's performance drop is even more pronounced and tends to classify more samples as fake.

Insight 4: *Alterations caused by common codecs such as Opus can negatively influence the performance and bias recognizes towards one class.*

4.7. Channel Losses

Channel losses are an unavoidable part of communication channels. Audio signals transmitted over communication channels are subject to the quality of the network and the characteristics of the medium. Fake audio detectors must be able to discern audio samples correctly even with channel losses. We consider downsampling the audio to simulate a low-bandwidth communication. For the experiment, we downsample the audio samples to 3.4 kHz before forwarding

DATASET	MODEL NAME	ACCURACY	TPR@ 0.05 FPR	AUC SCORE
BALANCED ASVSPOOF2021	RESNET-18	0.602 (0.727)	0.254 (0.517)	0.669 (0.754)
	RAWPC-DARTS	0.750 (0.796)	0.588 (0.630)	0.791 (0.802)
	RAWNET-2	0.543 (0.787)	0.142 (0.634)	0.618 (0.854)
	WHISPER FEATURES	0.501 (0.819)	0.205 (0.576)	0.748 (0.924)
OUT-OF-DISTRIBUTION	RESNET-18	0.441 (0.595)	0.062 (0.860)	0.404 (0.952)
	RAWPC-DARTS	0.642 (0.642)	0.320 (0.320)	0.686 (0.686)
	RAWNET-2	0.598 (0.598)	0.220 (0.220)	0.813 (0.813)
	WHISPER FEATURES	0.483 (0.411)	0.027 (0.036)	0.488 (0.382)

Table 6. Performance of selected detection models after downsampling to 3.4 kHz (Comparison to balanced ASVspoof2021 and OOD dataset in brackets respectively).

them to the detectors.

The results in Table 6 show the reduction in performance for the ResNet-18 model for both datasets. However, RawNet-2 and Whisper features particularly fail for the balanced dataset, resulting in a significantly lower TPR @ 0.05 FPR. Again, it has a shift toward classifying all audio as fake.

Insight 5: *Most models show difficulties for downsampled audio and detect input more likely as fake.*

5. Discussion and Conclusion

We conducted a comprehensive analysis of the performance of four generated audio detectors in real settings in which the audio input to these detectors is noisy, suffers from channel loss (downsampling), and alterations of their power/volume. In general, we observed that all tested classifers struggle for OOD samples such as new generative models but also alterations caused by typical communication channels. An overview of all the results for different alterations is shown in Appendix B in Table 7. The models tend to classify all samples as more likely to be fake or real, showing an insufficient generalization of the model. This is particularly evident for the sound level modifications, where we have shown that we can control the output of the classifier to be fake or real by just adjusting the sound level, independent of its input. This might be a result of the dataset from the ASVSpoof family which is, based on our literature review, the predominantly used dataset for training and testing of fake audio classifiers. Although the overall performance is low, we could also observe that models like the RawPC-Darts are less susceptible to channel effects.

We concluded that current detectors cannot reliably distinguish real and fake audio samples in realistic settings and that detectors for fake audio need to be improved for being used in detection in real-world environments.

References

Amazon.com, Inc. Amazon Polly: Deploy High-Quality, Natural-Sounding Human Voices in Dozens of Languages, October 2023. https://aws.amazon.com/polly/, as of July 26, 2024.

Ballesteros, D. M., Rodriguez-Ortega, Y., Renza, D., and Arce, G. Deep4SNet: Deep Learning for Fake Speech Classification. *Expert Systems with Applications*, 108: 115465:1–115465:12, December 2021.

Chen, H. and Magramo, K. Finance Worker Pays Out $25 Million after Video Call with Deepfake 'Chief Financial Officer', February 2024. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html, as of July 26, 2024.

Chen, L.-W. and Rudnicky, A. Fine-Grained Style Control In Transformer-Based Text-To-Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '22, pp. 7907–7911, Singapore, Singapore, May 2022. IEEE.

de Ruiter, A. The Distinct Wrong of Deepfakes. *Philosophy & Technology*, 34:1311–1332, June 2021.

ElevenLabs, Inc. ElevenLabs: Text to Speech & AI Voice Generator, October 2023. https://elevenlabs.io, as of July 26, 2024.

Ge, W., Patino, J., Todisco, M., and Evans, N. Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge Workshop*, ASVspoof '21, pp. 22–28, Virtual Conference, September 2021. ISCA.

Google, LLC. Text-to-Speech AI: Convert Text into Natural-Sounding Speech, October 2023. https://cloud.google.com/text-to-speech, as of July 26, 2024.

Huang, S.-F., Lin, C.-J., Liu, D.-R., Chen, Y.-C., and Lee, H.-y. Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 30:1558–1571, April 2022.

Jemine, C. M.Sc. Thesis: Automatic Multispeaker Voice Cloning – University of Liége, June 2019. https://matheo.uliege.be/handle/2268.2/6801, as of July 26, 2024.

Kawa, P., Plata, M., Czuba, M., Szymański, P., and Syga, P. Improved DeepFake Detection Using Whisper Features. In *Proc. INTERSPEECH 2023*, pp. 4009–4013, 2023. doi: 10.21437/Interspeech.2023-1537.

Kim, J., Kong, J., and Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *International Conference on Machine Learning*, ICML '21, pp. 5530–5540, Virtual Conference, July 2021. PMLR.

Li, Y. and Lyu, S. Exposing Deepfake Videos by Detecting Face Warping Artifacts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, CVPR '19, pp. 46–52, Long Beach, California, USA, June 2019. IEEE.

Liu, S., Su, D., and Yu, D. DiffGAN-TTS: High-Fidelity and Efficient Text-to-Speech with Denoising Diffusion GANs. *CoRR*, abs/2201.11972:1–16, January 2022.

Lukose, S. and Upadhya, S. S. Text to Speech Synthesizer-Formant Synthesis. In *International Conference on Nascent Technologies in the Engineering Field*, IC-NTE '17, pp. 1–4, Vashi, India, January 2017. IEEE.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvaturey. In *International Conference on Machine Learning*, ICML '23, pp. 202:24950–202:24962, Honolulu, Hawaii, USA, July 2023. PMLR.

Mozilla Foundation, and Community. Mozilla: Common Voice, June 2017. https://commonvoice.mozilla.org/cv/datasets, as of July 26, 2024.

Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., and Böttinger, K. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*, ICLR '21, pp. 1–15, Virtual Conference, May 2021. OpenReview.net.

Seitz-Wald, A. and Memoli, M. Fake Joe Biden Robocall Tells New Hampshire Democrats Not to Vote Tuesday, January 2024. https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984, as of July 26, 2024.

Tabet, Y. and Boughazi, M. Speech Synthesis Techniques. A Survey. In *International Workshop on Systems, Signal Processing and their Applications*, WOSSPA '21, pp. 67–70, Tipaza, Algeria, May 2021. IEEE.

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., and Larcher, A. End-to-End Anti-Spoofing with RawNet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '21, pp. 6369–6373, Toronto, Ontario, Canada, June 2021. IEEE.

Valin, J., Vos, K., and Terriberry, T. Definition of the Opus Audio Codec. RFC 6716, RFC Editor, September 2012. URL https://www.rfc-editor.org/info/rfc6716.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. CNN-Generated Images Are Surprisingly Easy to Spot… For Now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '20, pp. 8692–8701, Seattle, Washington, USA, June 2020. IEEE.

Wu, Y., Tan, X., Li, B., He, L., Zhao, S., Song, R., Qin, T., and Liu, T. AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios. In *Conference of the International Speech Communication Association*, INTERSPEECH '22, pp. 2568–2572, Incheon, Korea, September 2022. ISCA.

Xiph.Org, Foundation. Opus Interactive Audio Codec, September 2012. https://opus-codec.org, as of July 26, 2024.

Xue, J., Deng, Y., Han, Y., Li, Y., Sun, J., and Liang, J. ECAPA-TDNN for Multi-Speaker Text-to-Speech Synthesis. In *IEEE International Symposium on Chinese Spoken Language Processing*, ISCSLP '22, pp. 230–234, Singapore, Singapore, December 2022. IEEE.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

Zhang, Y., Jiang, F., and Duan, Z. One-Class Learning Towards Synthetic Voice Spoofing Detection. *IEEE Signal Processing Letters*, 28:937–941, April 2021.

A. Evaluating the Influence of the Amplitude

To properly observe the pattern traced by the reducing amplitude of the audio samples on the detection models, we perform an incremental decrease in amplitude/power, 1dB at a time. The resultant graph is shown in Figure 5.

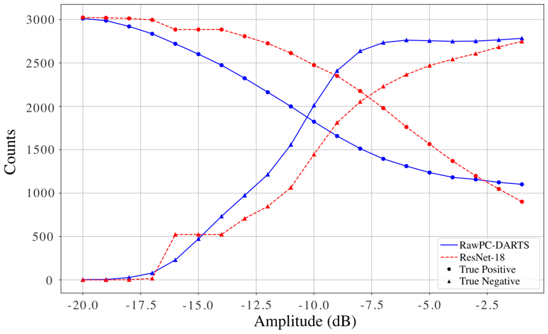

Figure 5. TPs and TNs for different amplitude levels over the OOD dataset for RawPC-DARTS and RawNet-2.

B. Comparison of Audio Alterations

Comparison of audio alterations caused by channel transmissions and transmissions over the air (Table 7).

Dataset	Manipulation	ResNet-18			RawPC-DARTS			RawNet-2			Whisper features			Average Accuracy
		Accuracy	TPR@5%FPR	AUC	Accuracy	TPR@5%FPR	AUC	Accuracy	TPR@5%FPR	AUC	Accuracy	TPR@5%FPR	AUC	
Balanced ASVspoof2021	Baseline	0.727	0.517	0.754	0.796	0.630	0.802	0.787	0.634	0.854	0.819	0.576	0.928	0.782
	Amplitude Reduction	0.715	0.503	0.740	0.721	0.498	0.784	0.732	0.417	0.781	0.797	0.417	0.873	0.741
	Opus codec	0.546	0.054	0.574	0.808	0.641	0.806	0.566	0.190	0.496	0.828	0.548	0.913	0.687
	Downsampling	0.602	0.254	0.669	0.750	0.588	0.791	0.543	0.142	0.618	0.501	0.205	0.748	0.599
Out-of-Distribution	Baseline	0.595	0.860	0.952	0.642	0.320	0.686	0.598	0.220	0.813	0.411	0.036	0.382	0.562
	Amplitude Reduction	0.525	0.698	0.946	0.508	0.334	0.649	0.536	0.062	0.483	0.427	0.028	0.375	0.499
	Opus codec	0.512	0.408	0.864	0.642	0.320	0.685	0.489	0.078	0.534	0.421	0.038	0.390	0.516
	Downsampling	0.441	0.062	0.404	0.642	0.320	0.686	0.598	0.220	0.813	0.483	0.027	0.488	0.541

Table 7. Performance of selected detection models after separately applying: 1) a 15 dB reduction in amplitude, 2) the Opus codec, and 3) downsampling to 3.4 kHz, compared to the baseline.